Modelação de um *Data Warehouse*

para a Direcção-Geral do Tesouro e Finanças

e implementação de um *Data Mart* para o processo de Gestão Patrimonial

por

Marco António Boialvo Gomes

Dissertação apresentada como requisito

parcial para a obtenção do grau de

Mestre em Estatística e Gestão da Informação

pelo

Instituto Superior de Estatística e Gestão da Informação

da

Universidade Nova de Lisboa

2010

à minha família.

Agradecimentos

Ao Prof. Dr. Miguel Neto pelo constante apoio e por acreditar na realização do presente trabalho.

Aos meus familiares, pela ajuda e incentivo, em especial à minha mulher.

À Direcção-Geral do Tesouro e Finanças pelo oportunidade realizar este trabalho.

A todos os outros, muitos, a quem devo o suficiente para ter conseguido efectuar a presente dissertação.

Resumo

Na Economia actual, as Tecnologias de Informação e do Conhecimento desempenham um papel cada vez mais importante no sucesso de Empresas e Governos. A complexidade dos dados apresentados a quem gere estas entidades torna a Gestão da Informação um aspecto fundamental a considerar. Assim, é necessário construir ferramentas que agreguem os dados, que os processem e que partilhem a informação obtida por toda a organização, permitindo a tomada de decisões mais rápidas e eficazes. De entre os Sistemas da Informação que se transformaram em elementos estratégicos para a Gestão, destacam-se o *Data Warehouse*, o *Data Mining* e a *Business Intelligence*. No presente trabalho procura-se demonstrar a importância destes Sistemas de Informação nas decisões da gestão, contribuindo de forma significativa para o processo de tomada de decisão. Evidencia-se essa importância apresentando um projecto de Sistema de Informação que faz uso da tecnologia *Data Warehouse* para a área de Gestão Patrimonial da Direcção-Geral do Tesouro e Finanças, para obter informação e conhecimento sobre a forma como são utilizados os imóveis do Estado.

Palavras-chave: *Data Warehouse*; *Data Mar*t; Gestão Patrimonial

Abstract

In today's Economy, Information Technologies play an increasingly important role in the success of Companies and Governments. The complexity of the data presented to those who manage these entities makes information management a key issue to consider. It is therefore necessary to build tools that aggregate data, process it and share the information obtained throughout the organization, thus allowing decision-making to became faster and more efficient. Among the Information Systems that have become strategic for Business Management some have stand out, such as Data Warehouse, Data Mining and Business Intelligence. This dissertation tries to demonstrate the importance of those Information Systems in management decisions, contributing to the process of decision making. That importance is outlined by presenting a project of an Information Systems that makes use of Data Warehouse technology in the area of Asset Management of the Direcção-Geral do Tesouro e Finanças, for acquiring information and knowledge about the use of State properties.

Key-words: *Data Warehouse*; *Data Mar*t; Asset Management

Índice

1 Introdução.....1
Prólogo.....3
Enquadramento.....5
Estrutura da dissertação.....6
2 Revisão da Literatura.....7
Data Warehouse.....9
O modelo Inmon.....11
Os três níveis de modelação.....12
A metodologia Meth2.....14
A Filosofia de Inmon.....15
O modelo Kimball.....16
Modelação dimensional.....16
Data Warehouse Bus.....17
Os quatro passos do processo de design.....18
A filosofia de Kimball.....20
Inmon vs Kimball.....21
3 Metodologia.....23
Motivação.....25
Metodologia.....25
Objectivos 26
4 Modelação do Data Warehouse.....27
A Direcção-Geral do Tesouro e Finanças.....29
Síntese histórica.....29
Missão.....30
Atribuições.....30
Os Sistemas de Informação da DGTF.....31
Regularizações e Recuperações Financeiras.....31
Intervenção Financeira do Estado.....31
Gestão Patrimonial.....31
Transversais.....32
Análise.....32
A modelação.....35

5 Implementação do Data Mart.....37
Modelação Dimensional.....39
Selecção do processo de negócio.....39
Definição da granularidade.....41
Escolha das dimensões.....43
Escolha dos factos.....44
Modelo Teórico.....45
Modelo Físico.....46
Definição do DWB.....53
Implementação Física.....54
Carregar o Data Mart.....55
ETL automático.....58
6 Resultados.....63
Do Data Warehouse.....65
Do Data Mart.....67
Relatórios.....68
7 Conclusão.....69
Discussão.....71
Considerações finais.....73
Referências.....75
Livros.....75
Legislação.....79
WEB.....81
ANEXOS.....83

Índice de Figuras

Figura 1: Relação entre os níveis do modelo de dados, (Inmon, 2002).....13
Figura 2 – O Meth2 (Inmon, 2002).....14
Figura 3 - Tabelas de Dimensão e de Factos – Vendas a retalho.....19
Figura 4: Sistemas de informação da DGTF.....34
Figura 5: Modelo de dados do SIIE.....42
Figura 6: Modelo teórico.....45
Figura 7: Modelo físico da AT.....52
Figura 8: Diagrama do DM em SQL Server 2005.....54
Figura 9: Pentaho Data Integration (Kettle).....55
Figura 10: Steps para Jobs.....56
Figura 11: Steps para Transformations.....56
Figura 12: ETL_SIIE.....58
Figura 13: O Job Proprietário.....58
Figura 14: Transformation E_T_Proprietario.....59
Figura 15: carrega AT_Quorum.....59
Figura 16: LookUp_Quorum.....60
Figura 17: LooUp_Proprietario.....61
Figura 18: Step Table output.....61
Figura 19: Carregamento de dados no Data Mart.....62
Figura 20: Actualização da tabela AT_Quorum.....62
Figura 21: A primeira iteração do DWB.....66
Figura 22: Modelo físico do DM.....67

Índice de Tabelas

Tabela 1 - Vantagens VS Desvantagens (Breslin, 2004).....21
Tabela 2: Nível de acesso por Perfil.....40
Tabela 3: Requisitos por perfil de acesso.....40
Tabela 4: Dimensões das tabelas de Factos.....43
Tabela 5: Atributos das tabelas de Factos.....44
Tabela 6: Matriz do Data Warehouse Bus.....53
Tabela 7: Matriz do DWB.....65

Siglas

TIC – Tecnologias de Informação e Comunicação

PN – Processos de Negócio

SI – Sistema de Informação

DW – *Data Warehouse*

ETL – *Extraction, Transformation and Loading*

SAD – Sistema de Apoio a Decisão

BD – Base de Dados

DGTF – Direcção-Geral do Tesouro e Finanças

SIIE – Sistema de Informação de Imóveis do Estado

RIAP – Recenseamento dos Imóveis da Administração Pública

MFAP – Ministério das Finanças e da Administração Pública

DM – *Data Mart*

DWB – *Data Warehouse Bus*

ERD – *Entity Relationship Diagram*

DIS – *Data Item Set*

MD – Modelação Dimensional

PGPI – Programa de Gestão e Inventariação do Património Imobiliário Público

RCM – Resolução de Conselho de Ministros

DSS – *Decision-Support System*

INE – Instituto Nacional de Estatística

RIGORE – Rede Integrada de Gestão dos Recursos do Estado

AT – Área de Trabalho

AP – Administração Pública

1 Introdução

Prólogo

Como resultado da evolução das Tecnologias de Informação e Comunicação (TIC), bem como do aumento da capacidade de processamento dos computadores, praticamente todas as empresas utilizam sistemas informáticos para suportar os seus processos de negócio (PN). Com o passar do tempo, estes sistemas acabam por gerar uma enorme quantidade de dados relacionados com o negócio (*Porter e Millar*, 1985). Estes dados, que estão armazenados nos sistemas transaccionais, são um recurso que, de uma forma geral, não é utilizado (*Subramanian, Smith e Nelson*, 1996). Efectivamente, os sistemas transaccionais não são projectados para produzir informações estratégicas, o que torna esses sistemas inapropriados para o apoio à tomada de decisões (*Gupta*, 1997).

A necessidade de obter informação estratégica, a partir que grandes volumes de dados dispersos por sistemas heterogéneos, levou a que fosse desenvolvido um novo género de Sistema de Informação (SI) designado de *Data Warehouse* (DW), estes SIs são construídos com o intuito de apoiar o processo de tomada de decisão na organização (*Boar*, 1997).

Segundo *Bill Inmon* (2002) o DW é definido como:

"*... a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions.*"

A definição de *Ralph Kimball* (2002), apresentada na obra *The Data Warehouse Toolkit,* define DW como:

"… a copy of transaction data specifically structured for query and analysis."

O DW não é uma aplicação que se compra e instala nos computadores da empresa. Na realidade, a sua implementação exige a integração de vários produtos e processos. Numa perspectiva minimalista, um DW não é mais que uma Base de Dados especializada, que integra e gere a recolha de informações a partir de sistemas transaccionais internos e fontes de dados externas. Neste contexto, um DW é construído para permitir: uma vista integrada e completa de toda a organização; acesso aos dados históricos da organização; ter uma fonte de dados verosímeis dentro da organização e facilitar os processos de tomada de decisão, sem sobrecarregar os sistemas operacionais (*Poe, Klauer e Brobst*, 1998).

Com o surgimento dos DWs foi necessário criar novos métodos de estruturação de dados, tanto no armazenamento como na consulta da informação. As empresas produzem e armazenam um volume elevado de dados, sendo normal que estes dados estejam dispersos por vários servidores, que podem, inclusive, estar dispersos por várias localizações geográficas e ter sido desenvolvidos em plataformas e linguagens diferentes (*Gupta*, 1997). Um dos desafios da implementação de um DW é a integração dos dados, eliminando as redundâncias, identificando os duplicados que possam estar em sistemas distintos, representadas sob formatos ou designações diferentes (*Adelman e Moss*, 2000). O processo de passagem dos dados dos sistemas transaccionais para o DW é denominado *Extraction, Transformation and Loading* (ETL), em Português, Extracção, Transformação e Carregamento. Inicialmente, os dados são extraídos para a *Staging Area*, em Português, Área de Trabalho (AT) e, depois de transformados e limpos, são carregados no DW. O sistema de DW, de uma forma geral é separado das bases de dados transaccionais, pelo que as consultas dos utilizadores ao mesmo não degradam a performance dos sistemas transaccionais, que ficam simultaneamente protegidos de alterações e perdas causadas pela manipulação indevida da informação.

Um DW oferece os meios necessários ao funcionamento de um Sistema de Apoio à Decisão (SAD) eficiente, fornecendo dados integrados e históricos, desde o topo da organização, a Administração, que necessita de informações mais resumidas e abrangentes, até ao nível mais baixo, o Operacional, onde os dados mais detalhados informam o enquadramento táctico de uma determinada área de negócio da empresa (*Shim, et al.* 2002). Deste modo, os utilizadores podem obter, de forma rápida, respostas às perguntas que os sistemas transaccionais não conseguem dar, permitindo tomar decisões com base em factos e não em intuições ou especulações (*Greenfield*, 2002).

Graças aos avanços na tecnologia de base de dados (BD) relacionais e dimensionais, ao processamento em paralelo, à tecnologia da computação distribuída e ao *software* livre, a criação de um DW está hoje ao alcance de todas as organizações. Estudar e conhecer a tecnologia de DW pode ajudar os empresários a descobrir novas formas de competir numa economia global, fazendo chegar ao mercado melhores produtos, de forma mais rápida que os concorrentes, sem aumento dos custos.

Não existem metodologias formais para implementação de um DW, ele deve ser adaptado às características e às necessidades de cada entidade, tendo como objectivo principal descobrir maneiras diferentes de actuar no mercado e averiguar quais as mudanças internas que devem ocorrer para a empresa se adaptar às mutações constantes da realidade, ou seja, utilizar o DW como fonte de inovação na gestão dos PNs.

Neste contexto, a qualidade dos dados é um factor importante. Uma má implementação dos sistemas transaccionais poderá originar grandes quantidades de dados incorrectos ou incoerentes. Tal facto torna a tarefa de extracção e limpeza de dados num factor de risco, que irá consumir tempo e recursos ao projecto, podendo inclusive atrasá-lo ao ponto de se tornar uma tarefa quase impossível (*Lee, Ling e Ko*, 1999; *Maletic e Marcus*, 2000).

Enquadramento

A ideia dos organismos do Estado pagarem valores associados aos rácios de ocupação dos imóveis vem de Governos anteriores. Para avaliar esta política foi efectuado um recenseamento dos imóveis do Estado no ano de 2004, conforme disposto na Resolução de Conselho de Ministros n.º 40/2004, de 29 de Março, denominado Recenseamento dos Imóveis da Administração Pública (RIAP). No ano de 2005, o Ministério das Finanças e da Administração Pública (MFAP) voltou a colocar a questão, tendo sido efectuada uma segunda tentativa de recenseamento com suporte legal na Resolução de Conselho de Ministros n.º 1/2006 de 2 de Janeiro, denominado RIAP 2 (INE, 2006). No entanto, este processo acabou por terminar sem se conseguir inventariar todos os imóveis.

No seguimento das políticas do XVI Governo Constitucional para a racionalização de recursos na Administração Pública, o Decreto-Lei n.º 280/2007, de 7 de Agosto, veio disciplinar o regime do património imobiliário público, tendo como pressupostos a eficiência e o bom aproveitamento dos recursos públicos e a sua conformidade com a actual organização do Estado. Assim, foi aprovado em Conselho de Ministros o Programa de Gestão do Património Imobiliário do Estado (PGPI), para o quadriénio 2009-2012, conforme descrito na Resolução de Conselho de Ministros n.º 162/2008, de 24 de Outubro. Foi com o objectivo essencial de assegurar o pleno conhecimento do património imobiliário público que a Portaria n.º 95/2009, de 29 de Janeiro, deu corpo ao programa de inventariação previsto no Artigo 114º do Decreto-Lei n.º 280/2007.

Esta medida visa aumentar a eficiência da utilização dos imóveis e será introduzida de forma gradual até 2012. Com este programa, pretende-se definir indicadores de ocupação e de custos de utilização, fomentar o planeamento das necessidades de cada serviço, bem como a calendarização de venda e arrendamento de imóveis.

A DGTF tem por missão assegurar a gestão integrada do património do Estado, assim como a intervenção em operações patrimoniais do sector público, nos termos da lei, sendo, desta forma, responsável pelo cumprimento da Resolução de Conselho de Ministros n.º 162/2008, de 24 de Outubro, nomeadamente a inventariação dos imóveis e a cobrança dos valores associados aos rácios de ocupação.

Estrutura da dissertação

A dissertação que agora se apresenta encontra-se dividida em sete capítulos. No presente capítulo é efectuada uma introdução ao conceito de DW, seguido de um enquadramento do conceito com a realidade da DGTF e a motivação do presente trabalho. São também definidos os objectivos do projecto e a estrutura da tese.

Capítulo 2 - Neste capítulo é efectuada uma análise do "estado da arte" baseado na revisão da literatura sobre DW.

Capítulo 3 - Apresenta a metodologia que será aplicada ao processo de modelação do DW e à implementação do DM.

Capítulo 4 - É efectuada a modelação do DW.

Capítulo 5 - É descrita a implementação do DM.

Capítulo 6 - São apresentados os resultados do projecto.

Capítulo 7 - São discutidos os resultados e apresentadas as conclusões finais.

2 Revisão da Literatura

Data Warehouse

O DW é um grande repositório de dados, em forma de série temporal, usado para apoio à decisão. Os DWs contêm frequentemente volumes de dados na ordem dos *Terabytes,* que podem ser pesquisados e agregados pelos utilizadores. A maioria dos dados de um DW é constituída por transacções provenientes dos sistemas operacionais. Utilizando software especializado é possível extrair os dados das BDs operacionais, processar, sintetizar e agregar os mesmos, para que sejam armazenados no DW numa forma mais eficiente para a sua exploração.

Uma organização tem à sua escolha um grande conjunto de ferramentas de design, armazenamento e manutenção, comerciais ou de *software* livre, para implementar o DW. Nem todas as ferramentas são compatíveis entre si e nem todas são apropriadas para a metodologia de desenvolvimento escolhida. Apesar do grande universo de escolha nas ferramentas, a modelação das mesmas está geralmente baseada numa de duas metodologias: *Inmon* ou *Kimball*.

Escolher entre *Inmon ou Kimball* corresponde a uma escolha a nível de arquitectura e metodologia. Entender os fundamentos da arquitectura e da metodologia dos dois modelos proporciona um conhecimento básico do que é, e como funciona, um DW.

Em 1990, *Bill Inmon* ganhou o apelido "Pai do *Data Warehouse*" apresentando o termo *Data Warehouse* na publicação *Building the Data Warehouse*. As empresas começaram, desde então a implementar a visão de *Inmon*, com graus variados de sucesso. *Inmon* (1994, 1997) tem vindo a apresentar a sua visão sobre a metodologia a adoptar no desenvolvimento de DWs. Na terceira edição do seu trabalho (*Inmon,* 2002) descreve uma arquitectura lógica para extrair os dados de BDs operacionais dispersas. Os dados são transformados e organizados temporalmente numa única BD (*Data Warehouse*). Parte destes dados são então extraídos para BDs menores, criando BDs departamentais denominadas *Data Mart* (DM) onde os utilizadores finais exploram os dados e criam relatórios. Para criar o DW e os DMs, *Inmon* propõe uma metodologia *top-down*, partindo do geral para a pormenorização, dos vários sistemas que o compõem.

Depois da publicação do livro de *Inmon*, outros especialistas de BD começaram a criar DWs. A experiência de *Ralph Kimball* conduziu-o ao desenvolvimento de uma metodologia própria tendo, em 1998, publicado *The Data Warehouse Toolkit*. Depois de vários anos de experiência, *Kimball* (2002) publicou uma segunda edição da sua obra, recomendando nesta versão uma arquitectura de múltiplas BDs e DMs, organizadas por áreas de negócio*,* em que os DMs têm

que aderir a um canal de comunicação comum denominado *Data Warehouse Bus* (DWB). Nesta versão, o DW é definindo como sendo a soma dos vários DMs. Para o desenvolvimento é recomenda uma metodologia inversa à de *Inmon*, uma aproximação *bottom-up*, que parte da análise dos vários sistemas individuais terminando com a agregação dos mesmos num grande DW.

O DW existe para facilitar e apoiar o processo de decisão dentro das organizações. Os Sistemas de Apoio à Decisão (SAD) ajudam os utilizadores com análises *ad hoc*, relatórios automáticos e *dashboards*. Geralmente os SAD requerem dados históricos ao nível da transacção, no entanto, é possível que estes sejam agregados, permitindo aos utilizadores analisar facilmente grandes quantidades de dados. Para encontrar uma relação estatisticamente significativa entre itens num DW é necessário que os dados em análise sejam suficientemente detalhados para conter os dados da transacção e a descrição dos itens nela envolvidos. Isto evidencia porque os DWs contêm grandes quantidades de dados (Chaudhuri e Dayal, 1996) .

Uma exigência menos óbvia é poder observar os dados sem restrições prévias, ou seja, submeter questões em que podemos cruzar todos os dados disponíveis das fontes mais variadas (Hammer J., Garcia-Molina H., Jennifer W., Labio W. e Zhuge Y., 1995). Isto implica duas coisas, o DW tem que ser de fácil utilização e intuitivo para os utilizadores, e os tempos de resposta devem ser enquadrados num intervalo razoável para o utilizador não se dispersar noutras actividades.

O modelo Inmon

A arquitectura proposta por *Inmon* contém todos os SIs e BDs da organização. A este conjunto *Inmon* chama *Corporate Information Factory* (CIF) (Inmon, Imhoff e Sousa 1998). O autor divide o ambiente de BDs global da organização em quatro níveis:

1. Operacional
2. *Data Warehouse*
3. Departamental
4. Individual

O primeiro nível contém dados dos sistemas transaccionais. Este nível regista a actividade quotidiana da organização. Os dados são então extraídos desses sistemas, transformados e carregados no *Data Warehouse* (*Inmon*, 2002), correspondente ao nível dois.

Os dados contidos ao nível departamental (3) são resumos do DW e, dependendo das exigências de informação, assim será a agregação dos dados. Por exemplo: o departamento de crédito pode resumir os dados, considerando a informação do endereço de cliente como irrelevante, mantendo apenas um indicador, para a mudança de endereço. Por outro lado, o departamento de marketing poderia resumir mais os dados, deixando para trás todos os dados de contacto do cliente com excepção do código postal. A BD de cada departamento contém dados resumidos de acordo com suas necessidades. A arquitectura de *Inmon* assegura que todos os dados são consistentes porque todos os dados departamentais provêm do mesmo DW.

Os utilizadores individuais estão no quarto nível da arquitectura, quando estes separam conjuntos *ad hoc* de dados como parte da análise necessária ao suporte das tomadas de decisão. Este quarto nível é temporário e reside no computador pessoal do utilizador (*Inmon*, 2002). Como exemplo, podemos tomar um utilizador que trabalha no departamento de crédito; este pode seleccionar todas as contas que tiveram créditos faltosos, pelo menos uma vez, nos últimos três anos, para efectuar uma análise de risco das várias agências.

Se a BD do departamento não reteve os dados com o nível de detalhe necessário, é sempre possível analisar o DW e efectuar um novo carregamento da BD departamental com os detalhes omissos na versão anterior. O esforço inicial para construir o DW é útil e válido, porque permite a criação de uma infinidade de BD departamentais, sem correr o risco de ter dados incompatíveis ou duplicados (*Inmon*, 2002).

Os três níveis de modelação

Para a implementação do DW, *Inmon* propõe três níveis de modelação de dados. Seguindo o sistema de análise *top-down*, o primeiro nível designa-se *Entity Relationship Diagram* (ERD), em português, Diagrama Entidade-Relação. A equipa de desenvolvimento cria um ERD, à semelhança do que faria para um BD transaccional, para cada departamento que vai usar o DW. O ERD do DW é o resultado da integração de todos os ERD departamentais, contidos no âmbito do projecto (*Inmon,* 2002).

O segundo nível do modelação estabelece o modelo *Data Item Set* (DIS), para cada entidade do modelo ERD do departamento. A integração dos vários DIS departamentais compõe o DIS global do DW. O modelo de dados deste nível tem quatro construtores:

- Agrupamento primário de dados – Existe apenas uma vez para cada entidade do modelo ERD da DW. Possui atributos que só existem uma vez para cada entidade do modelo ERD, chaves que o relacionam com outros agrupamentos e Tipos de dados.
- Agrupamento secundário de dados – Contém atributos que podem existir várias vezes para cada entidade do modelo ERD. Deve haver tantos agrupamentos secundários, como grupos de dados que se repetem.
- Conector – Representa a relação entre entidades do modelo ERD.
- Tipo de dados – Contém os tipos de dados que podem ser gerados na entidade do modelo ERD.

Assim, o ERD criado no primeiro nível é a base para o DIS no nível intermédio. A figura 1 ilustra a relação ERD-DIS para cada visão do utilizador, mostra também como as várias visões dos utilizadores são combinadas num único ERD e DIS. Dentro de um DIS, cada rectângulo representa uma tabela lógica dentro de um departamento ou do DIS Global. Os rectângulos à direita do DIS representam o agrupamento secundário de dados (*Inmon*, 2002).

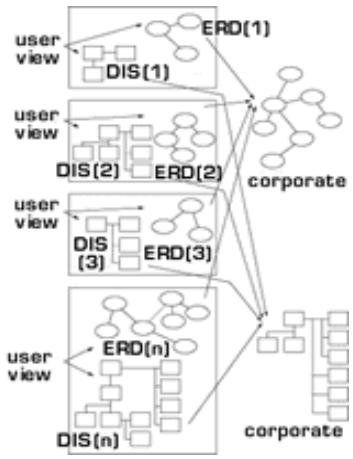

Figura 1: Relação entre os níveis do modelo de dados,
(Inmon, 2002)

Inmon usa um exemplo para ilustrar a diferença entre dados operacionais e dados armazenados no DW. Neste exemplo, a entidade é o cliente de um banco e o atributo em apreciação é a capacidade de crédito. A BD do sistema transaccional contém a avaliação de crédito actual do cliente e informação com interesse (empréstimos, contactos, saldo médio, etc.) num único registo. O DW contém a história de crédito para este cliente, agregada por ano, com um registo anual (*Inmon*, 2002).

No exemplo referido, a entidade cliente gera um agrupamento primário de dados, incluindo: contacto, conta, cartão, etc. O Item Conta tem várias iterações: à ordem, a prazo, poupança, criando desta forma o agrupamento secundário. Os conectores servem para relacionar o cliente com os vários tipos de conta que este possui. O quarto construtor serve para classificar os dados produzidos pelos vários itens do agrupamento secundário, permitindo categorizar os vários tipos de movimentos bancários associados à conta à ordem: debito, credito, transferência, etc.

Para criar os ERD e DIS departamentais e globais da Figura 1 é necessária uma grande capacidade de modelação relacional, bem como um conhecimento profundo dos processos de negócio. *Inmon* (2002) sugere o uso de *templates* de modelos de dados que já existem para vários PNs de modo a poupar tempo de desenvolvimento.

O nível mais baixo do modelo de dados é o físico. Nas palavras de *Inmon* (2002):

"…The physical data model is created from the midlevel data model merely by extending the midlevel data model to include keys and physical characteristics of the model. "

Quando os três níveis da modelação estão completos, o desenvolvimento da DW pode começar.

A metodologia Meth2

Completar os três níveis de modelação é condição prévia para usar a metodologia de *Inmon*, esta não é mais que uma adaptação do desenvolvimento em espiral de *Boehm* (1988), que este denominou *Meth2, sendo Meth1 p*ara desenvolver Sistemas Transaccionais e *Meth3 p*ara optimização de DWs existentes). *Inmon* designou o primeiro passo da modelação de três níveis, *Decision Support 1* (DSS1). A metodologia é constituída por dez passos, representados na Figura 2.

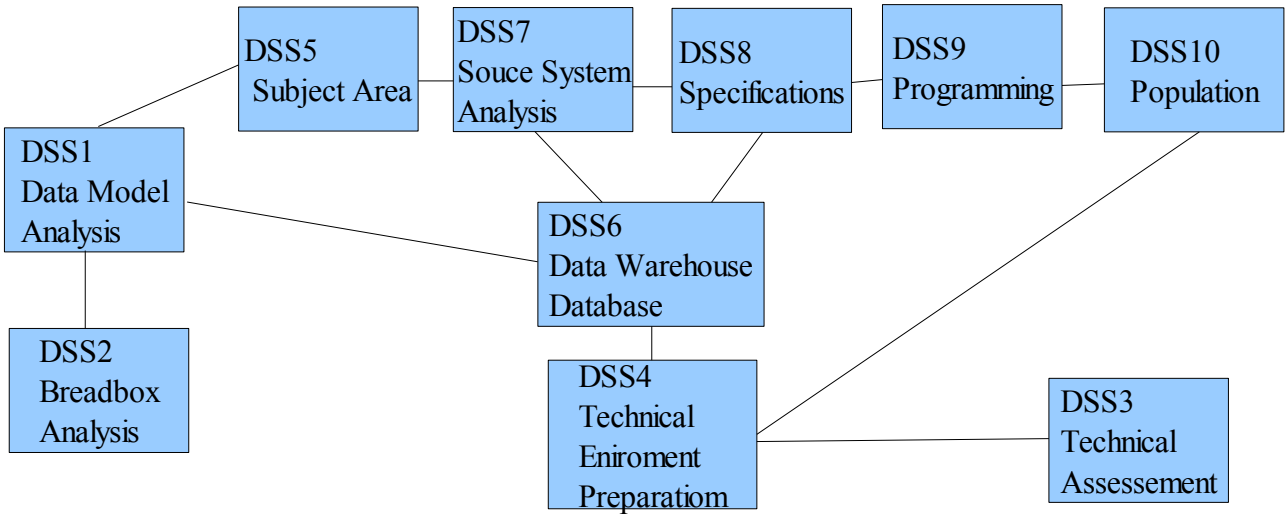

Figura 2 – O Meth2 (Inmon, 2002).

Usando o modelo de três níveis de dados como ponto de partida, o próximo passo é a análise de granularidade (DSS2) designada por *Breadbox Analysis*. A granularidade é a medida do detalhe dos dados. Os dados transaccionais têm o nível de granularidade mais baixo, porque têm a informação mais detalhada possível, este nível tem também a designação de atómico. Se o volume de dados for muito grande, então a equipa de desenvolvimento deve considerar diferentes níveis de granularidade para os dados (*Inmon*, 2002). Isto envolve armazenar alguns dados ao nível da transacção e outros em formas resumidas (ex: total diário, vendas anuais).

Quando a questão da granularidade estiver resolvida, é seleccionada a primeira área de negócio (DSS5). Esta vai tornar-se a primeira BD departamental a ser criada. A equipa de desenvolvimento analisa os sistemas transaccionais que fornecem os dados da primeira área de negócio (DSS7), definem as especificações (DSS8), programam as especificações (DSS9) e carregam a BD (DSS10). Entretanto, o DW global começa a ser desenvolvido paralelamente (DSS6) e, quando houver informação suficiente é efectuada uma avaliação técnica (DSS3).

Esta avaliação assegura que os dados na DW estão acessíveis e bem administrados (*Inmon*, 2002).

Com o desenvolvimento, sucessivo, das BD departamentais, o DW é afectado pelas alterações. Na Figura 2 podemos ver o aspecto repetitivo das alterações, representado pelas linhas que ligam os vários passos ao DW. As linhas ligam a análise dos sistemas operacionais (DSS7) e as especificações (DSS8) com o DW (DSS6). Isto significa que o DW será revisitado cada vez que uma BD departamental nova é desenvolvida. A linha que liga o carregamento da BD departamental (DSS10) com a preparação do ambiente técnico (DSS4) demonstra a natureza repetitiva de Meth2. Com a preparação do ambiente técnico (DSS4), *Inmon* refere-se à verificação dos recursos de rede, significando que *hardware* de armazenamento, sistema operativo e software de acesso estão prontos para receber dados (*Inmon*, 2002).

A metodologia de desenvolvimento em espiral de *Inmon* é orientada à informação:

"Um dos aspectos mais salientes da metodologia orientada para a informação é que assenta em esforços anteriores, utilizando código e processos anteriormente desenvolvidos." (*Inmon*, 2002)

O modelo de três níveis assenta na metodologia em espiral, impondo que a perspectiva dos utilizadores seja consistente com o modelo global. O processo desenvolve as BD departamentais, subsequentes, usando o código e processos que serviram para criar as BD departamentais anteriores. Isto diminui o tempo e esforço necessários para a construção das BD seguintes, diminuindo consideravelmente o tempo necessário para percorrer as etapas do processo, de DSS1 até DSS10.

A Filosofia de Inmon

Inmon vê o DW como uma parte integrante da CIF (Inmon et al. 1998); isto significa que o DW e as BDs transaccionais são parte de um todo maior. A aproximação evolutiva de *Inmon* tem por base a tecnologia de BDs relacionais, os seus métodos de desenvolvimento o design de sistemas transaccionais, assentando em princípios e práticas usadas nas décadas anteriores. Podemos assim afirma neste contexto que o modelo de *Inmon* surge como uma evolução natural do modelo relacional.

Por outro lado, uma consequência da metodologia de *Inmon* é os seus principais destinatários serem profissionais das TIC, pois é necessário um grande conhecimento das ferramentas e metodologias de análise e desenvolvimento, isto implica que os utilizadores finais tenham papéis passivos ou acessórios no desenvolvimento do DW, podendo desta forma diminuir a aceitação do projecto (*Breslin*, 2004).

O modelo Kimball

O modelo de *Kimball* difere em muitos aspectos da aproximação tradicional das BDs relacionais. Uma diferença significativa é que os DWs construídos com o modelo *Kimball* usam uma nova metodologia designada de Modelação Dimensional (MD). Outra diferença importante é a arquitectura global ser caracterizada por múltiplas BDs, designadas *Data Marts,* que são interligadas entre si pelo *Data Warehouse Bus,* responsável pela coerência da informação integrante do DW.

Modelação dimensional

A MD pode parecer estranha aos profissionais das TIC, familiarizados com a modelação relacional, pois a MD começa com tabelas em vez de ERD. Estas tabelas podem ser de dois tipos, Factos ou Dimensões. As tabelas de Factos contêm valores, enquanto as tabelas de Dimensão contêm medidas das valores contidos nas tabelas de Factos. As tabelas de Factos contêm grupos repetidos, o que viola as regras da normalização do modelo relacional. Contudo, esta violação das regras da normalização tem como objectivo aumentar o desempenho do DW.

No livro de *Kimball* (2002) é apresentado um exemplo que ajuda a compreender a modelação dimensional, do qual nos vamos socorrer para explicar a MD. Na modelação de um DW para um supermercado temos uma tabela de factos denominada Total_Vendas_Diario. Esta tabela contém cinco colunas: produto, loja, data, quantidade vendida e o valor das vendas em Dólar. As tabelas de Dimensão neste exemplo são Tempo, Loja e Produto.

As tabelas de Factos contêm muitas linhas e poucas colunas. Isto é essencial para facilitar o uso e aumentar o desempenho. O número de linhas na tabela Total_Vendas_Diario pode ser calculada através do número de produtos vendidos por dia em cada loja. *Kimball* estima que a tabela Total_Vendas_Diario contenha milhões de linhas e ocupe aproximadamente 10 GB, em apenas cinco colunas. Assim, basta adicionar uma coluna para o tamanho da tabela aumentar 2 GB (*Kimball*, 2002). Este exemplo ilustra a importância de manter o número de colunas tão baixo quanto possível.

Pelo contrário, é possível que as tabelas de Dimensão contenham centenas ou milhares de linhas e terem centenas de colunas, correspondentes a alguns Megabytes, porque estas tabelas contêm todos os atributos das valores da tabela de Factos de forma desnormalizada. A chave primária da tabela da Dimensão Produto é a chave de produto. O resto das colunas desta tabela de dimensão são atributos do produto, tais como: descrição do produto, marca, descrição do tipo de pacote, descrição do departamento, peso, etc.

A Dimensão Tempo e a Dimensão Loja também têm grande número de colunas, mas relativamente poucas linhas (*Kimball*, 2002).

Em termos de funcionamento *Kimball* (2002) afirma o seguinte:

"...A database engine can make very strong assumptions about first constraining the heavily indexed dimension tables, and then attacking the fact table all at once with the Cartesian product of the dimension table keys satisfying the user's constraints."

A MD é uma abordagem que capitaliza na estrutura única do DW, sendo essencial manter as tabelas de Factos com poucas colunas e desnormalizar as tabelas de Dimensão para que tenha muitas colunas e poucas linhas. O *Data Mart* resultante é fácil de utilizar pelo utilizador final e permite tempos de resposta curtos, porque os critérios para resumir os dados já estão nas tabelas de Dimensão. Esta característica vai de encontro a um dos principais objectivos da construção de um DW: a facilidade de uso e de acesso à informação.

Data Warehouse Bus

Na arquitectura de *Kimball*, os dados são copiados dos sistemas operacionais para uma área de trabalho onde são tratados e depois carregados em DMs. Estas são as fontes de dados para as pesquisas dos utilizadores finais.

Tipicamente cada DM está baseado num único processo de negócio. Alguns exemplos de processos de negócio são: vendas, inventário (stocks), aluguer. Mais que um departamento pode ter interesse num processo de negócio, assim, nenhum departamento é considerado o dono do DM (*Kimball*, 2002).

O DWB é a charneira dos DMs da arquitectura de *Kimball*, permitindo que a soma dos DMs possa funcionar como um todo. A arquitectura do DWB é outra forma de dizer que todos os DMs têm que usar dimensões *standard*. A única exigência de uma dimensão standard é que as chaves, nomes de coluna, definições de atributos e valores de atributo sejam consistentes ao longo dos vários DM, e por conseguinte, dos vários PNs. Posto de outra forma, duas dimensões são *standard* quando são coincidentes, ou uma é um subconjunto da outra. Mais importantes, os nomes das colunas produzidos em resposta a pesquisas devem ser coincidentes (*Kimball*, 2002). Este conjunto complicado de exigências é cumprido com o auxilio da MD e a observância dos quatro passos do processo de design, que se apresentam em seguida.

Os quatro passos do processo de design

Kimball apresenta uma metodologia própria para o desenvolvimento de DWs. Esta envolve uma aproximação *bottom-up*, consistindo na construção de um DM de cada vez, apoiado no DWB.

Os quatro passos do processo de design são:

- Seleccionar o processo de negócio;
- Definir o grão;
- Escolher as dimensões;
- Identificar as métricas da tabelas de factos.

A definição de PN é bastante ampla. Por exemplo, o processamento de ordens de compra é um PN do interesse do departamento de vendas, marketing, finanças e inventário. Para escolher o primeiro processo de negócio para integrar no DW, deve ser seleccionado o processo com maior impacte na instituição (*Kimball*, 2002).

Definir a granularidade é o processo de escolha do detalhe dos dados contidos no DW. O nível mais baixo de granularidade é designado de atómico, significando que não pode ser dividido. Escolher um nível de grão atómico é desejável, porque permite aos utilizadores agregar os dados à sua vontade, enquanto que escolher um nível intermédio de grão implica o risco de não ser possível satisfazer todas as pesquisas solicitadas pelos utilizadores. Nas palavras de *Kimball* (2002):

"*Preferably you should develop dimensional models for the most atomic information captured by a business process...*"

No exemplo do supermercado, apresentado anteriormente, o grão escolhido será a linha de uma transacção do ponto de venda. As possibilidades de análise de dados com este nível de granularidade são ilimitadas. A granularidade ao nível atómico fornece apoio à tomada de decisão em virtualmente todos os aspectos da venda a retalho. Podemos incluir na lista de exemplos a avaliação de promoções, expansão de linhas de produtos e a redução das vendas de um produto às custas da promoção de outro.

Após a escolha da granularidade, o próximo passo é seleccionar as dimensões. No exemplo da venda a retalho, as dimensões são o Tempo, a Loja e o Produto. Cada uma das tabelas de dimensão tem um grande número de atributos. Ao contrário do que defendem os analista de modelos relacionais, a tabela da dimensão Tempo inclui muitos atributos, tais como: número do dia da semana, número do dia do mês, número do dia do ano, Número de Semana,

Número de Mês e assim por diante. *Kimball* (2002) justifica esta tabela desnormalizada, mostrando que os valores de dez anos destes dados geram aproximadamente 3650 linhas resultando num ficheiro com apenas dezenas de *kilobytes*.

O quarto passo é determinar quais as métricas que devem ser incluídas nas tabelas de factos. No exemplo da venda a retalho, são incluídos alguns valores calculados junto com os valores atómicos, facilitando as pesquisas ao utilizador ao mesmo tempo que aumentam o desempenho do DW. Os valores na tabela de Factos da vendas de retalho são: tempo, produto, loja, número de transacção, quantidade de vendas, quantidade de vendas em Dólar, quantidade de custos em Dólar, lucro em Dólar. Incluir o lucro em Dólar é um exemplo de melhoria desempenho que é conseguida à custa de violar as regras das Modelação Relacional. Os utilizadores interrogam frequentemente o DW pelo lucro total; assim, incluir este valor calculado na tabela de Factos melhora o desempenho da pesquisa.

O resultado dos quatro passos é mostrado de forma simplificada na Figura 3. A tabela de Factos é apresentada com todos os factos, mas nas tabelas de Dimensão só aparecem as chaves primárias. Cada uma das tabelas de Dimensão mostradas na figura 3 tem dezenas de atributos, que permite aos utilizadores compor um conjunto bastante rico de pesquisas.

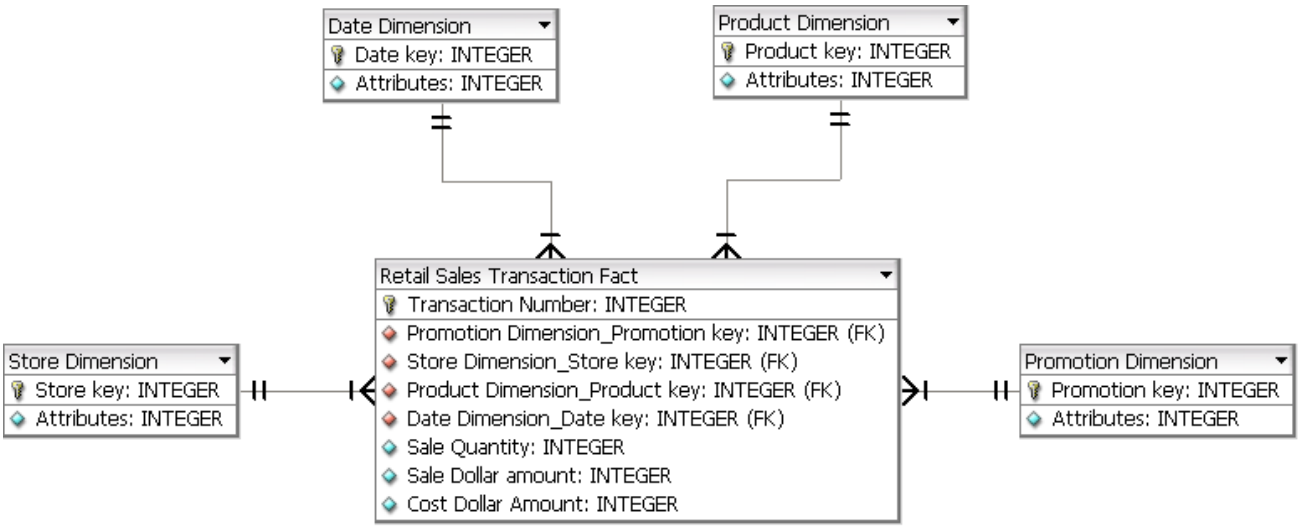

Figura 3 - Tabelas de Dimensão e de Factos – Vendas a retalho

A filosofia de *Kimball*

As exigências empresariais orientam o processo e a natureza do DW. *Kimball* (2002) define as metas de um DW da seguinte forma:

- Facilidade de acesso à informação;
- Informação organizada de forma consistente;
- Resistente à mudança;
- Suporte ao processo de decisão.

A lista termina com uma advertência disfarçada de objectivo:

"*The business community must accept the data warehouse if it is to be deemed successful*".

Para *Kimball*, a aceitação do DW é medida pela sua utilização, como indicador da sua simplicidade de utilização. A metodologia de desenvolvimento de *Kimball* é suficientemente fácil para o utilizador final participar activamente (*Breslin*, 2004). O exemplo da venda a retalho, da Figura 3, exibe a facilidade de uso da DM, tanto os atributos como as relações entre as tabelas de Factos e Dimensão são familiares aos utilizadores que necessitam de pesquisar os dados de vendas de retalho.

Inmon vs Kimball

Na tabela 1 podemos comparar de forma resumida os pontos fortes e as desvantagens de cada um dos modelos de implementação de DWs.

	Vantagens	Desvantagens
Inmon	Engloba todos os SI da empresa Construído de raiz Centralizado	Demorado Tecnicamente complexo
Kimball	Resultados rápidos Pouco complexo	Descentralizado Perigo de ficar a meio

Tabela 1 - Vantagens VS Desvantagens (*Breslin*, 2004).

Para *Bill Inmon* o DW é uma parte do sistema global de *BusinessIintelligence*. Uma empresa tem um DW, e os DM obtêm a sua informação do DW. A informação é guardada no DW na 3ª forma normal.

Ralph Kimball considera o DW como sendo o conglomerado de todos os DMs da empresa. Informação é armazenado no modelo dimensional.

3 Metodologia

Motivação

A modelação de um DW para a centralização da informação das diversas áreas de actuação da DGTF, em especial para o PN da Gestão Patrimonial seriam uma mais-valia para a instituição. Este repositório de dados será o suporte de um conjunto de ferramentas de *Data Mining, Business Analytics* e *Reporting.* Estes instrumentos serão imprescindíveis para uma análise cuidada dos dados, permitindo um acompanhamento e avaliação mais rigorosos.

Metodologia

A estrutura da DGTF é constituída por três áreas principais:

1) Regularizações e Recuperações Financeiras, composta pela Direcção de Serviços de Regularizações Financeiras e a Direcção de Serviços de Gestão de Recursos.
2) Intervenção Financeira do Estado que é composta pela Direcção de Serviços de Apoios Financeiros, Direcção de Serviços de Gestão Financeira e Orçamental e a Direcção de Serviços de Participações do Estado.
3) E, finalmente, a Gestão Patrimonial que se divide na Direcção de Serviços de Apoio Técnico Patrimonial e a Direcção de Serviços de Gestão Patrimonial.

A estrutura reflecte a divisão entre áreas e as diferentes atribuições de cada uma, com cada área a perseguir objectivos distintos que apenas se cruzam através das unidades que têm um comportamento transversal à instituição (Contabilidade, Recursos Humanos e Informática).

Perante este cenário, a metodologia escolhida para a implementação do DW da DGTF foi a apresentada por *Kimball* (2002), pois permite actuar individualmente sobre cada área de negócio e apresentar rapidamente resultados sem que toda a instituição esteja coberta pelo projecto e a salvo da introdução de novos SIs ou alterações na orgânica da DGTF. Esta metodologia vai também permitir que o desenvolvido tenha intervenção directa dos colaboradores da área em análise, facilitando a sua aceitação.

A metodologia que será utilizada é a apresentada por *Kimball* (2002) com as respectivas adaptações ao caso da DGTF. Assim temos:

1 Definição das dimensões *standard* para o DWB;

2 Modelação Dimensional;

2.1 Seleccionar o processo de negócio;

2.2 Definir o grão;

2.3 Escolher as dimensões;

2.4 Identificar os factos.

Objectivos

O presente projecto tem dois objectivos:

- Um objectivo teórico que visa a modelação da um DW para a informação da DGTF, segundo a metodologia de *Kimball* (2002), enquanto instrumento agregador das várias áreas de negócio e normalizador da informação existente.

- Um objectivo prático, na sequência do anterior, que implementa um DM para um processo de negócio da DGTF.

4 Modelação do *Data Warehouse*

A Direcção-Geral do Tesouro e Finanças

A actual DGTF é um organismo com uma longa história na administração pública, a sua organização foi evoluindo no sentido de acompanhar as alterações estruturais e tecnológicas de quase três séculos. A sua criação remonta à Monarquia, tendo sido criado o Tesouro Público em 16 de Maio de 1833, em substituição do Erário Régio - "Tesouro do Soberano ou do Príncipe".

Síntese histórica

O Tesouro Público Nacional foi criado pelo Decreto n.º 22, de 16 de Maio de 1833, substituindo o Erário Régio. Com este Decreto foi efectuada a separação dos bens da Nação dos haveres do Soberano, estabelecendo-se o Tesouro Nacional como "a união de todos os direitos, rendas e bens da Fazenda Pública" (DGTF, 2009).

Com a implantação da I República, foi alterada a organização da Administração Financeira, tendo sido criada, em 14 de Janeiro de 1911, a Direcção-Geral da Fazenda Pública, cujo Director-Geral acumulava o cargo de Secretário-Geral do Ministério das Finanças. Este organismo integrou as competências das Direcções-Gerais da Tesouraria e dos Próprios do Estado e das Inspecções-Gerais do Tesouro e do Cadastro. Mais de meio século depois, o Decreto-Lei n.º 49-B/76, de 20 de Janeiro, divide esta Direcção-Geral, dando lugar à Direcção-Geral do Tesouro e à Direcção-Geral do Património.

A Lei Orgânica do MFAP, aprovada pelo Decreto–Lei n.º 158/96, de 3 de Setembro, alterou a estrutura da DGT, tendo transitado para novos serviços algumas das suas competências, salientando-se a criação do Instituto de Gestão do Crédito Público com o objectivo de gerir a dívida pública directa e o financiamento do Estado.

No actual quadro das orientações definidas pelo Programa de Reestruturação da Administração Central do Estado, a nova Lei Orgânica do MFAP, aprovada pelo Decreto-Lei n.º 205/2006, de 27 de Outubro, consagrou a reestruturação da Direcção-Geral do Tesouro, passando a denominar-se Direcção-Geral do Tesouro e Finanças.

Nesta sequência, o Decreto Regulamentar n.º 21/2007, de 29 de Março, aprovou a orgânica da DGTF, congregando parte das atribuições anteriormente prosseguidas pela DGP e pela Direcção-Geral dos Assuntos Europeus e Relações Internacionais e procedendo à transferência para o Instituto de Gestão do Crédito Público da gestão das disponibilidades de tesouraria (DGTF, 2009).

Missão

A DGTF tem por missão assegurar a efectivação das operações de intervenção financeira do Estado, acompanhar as matérias respeitantes ao exercício da tutela financeira do sector público administrativo e empresarial, da função accionista, e assegurar a gestão integrada do património do Estado, bem como a intervenção em operações patrimoniais do sector público (DGTF, 2009).

Atribuições

Para a prossecução da sua missão, a DGTF tem como atribuições:

- Controlar a emissão e circulação da moeda metálica;
- Administrar a carteira de participações do Estado;
- Assegurar o estudo, acompanhamento e intervenção nas matérias respeitantes ao exercício da tutela financeira do sector público, administrativo e empresarial e ao exercício da função accionista do Estado, nos planos interno e internacional, bem como nas matérias respeitantes ao acompanhamento das parcerias público-privadas e das concessões;
- Conceder subsídios, indemnizações compensatórias e bonificações de juros, nos termos previstos na lei e avaliar os resultados da política de apoios financeiros do Estado;
- Efectuar e controlar as operações activas, a nível interno e internacional;
- Assegurar a condução do processo de concessão de garantias do Estado e administrar a dívida pública acessória;
- Assegurar a gestão financeira de patrimónios autónomos;
- Dar apoio técnico à participação portuguesa nos assuntos relacionados com a União Económica e Monetária e assegurar a representação técnica do MFAP em organizações europeias e internacionais em matéria financeira, sem prejuízo das atribuições de orientação geral e estratégica do Gabinete de Planeamento, Estratégia, Avaliação e Relações Internacionais do MFAP;
- Adquirir, arrendar, administrar e alienar, directa ou indirectamente, os activos patrimoniais do Estado, bem como intervir, nos termos da lei, em actos de gestão de bens;

- Promover a recuperação de créditos do Tesouro;
- Assegurar a assunção de passivos de entidades ou organismos do sector público e a regularização de responsabilidades financeiras do Estado (DGTF, 2009).

Os Sistemas de Informação da DGTF

A DGTF tem um grande variedade de sistemas de informação com funções muito diversas. Neste ponto iremos analisar esses sistemas e o seu contributo para o DW.

Nem todos os sistemas podem ser integrados no DW, pois alguns são de origem exterior à DGTF, designadamente de privados ou de outros organismos da Administração Pública e, por outro lado, alguns destes contêm informação considerada sensível, pelo que o acesso foi condicionado ou negado.

Na seguinte lista, temos os SIs da DGTF organizados por área de actuação.

Regularizações e Recuperações Financeiras

INVESDOC - Aplicação da biblioteca para gestão de publicações e documentos.

CAE – Aplicação para gestão do Credito Agricula de Emergência.

SIRC – Sistema de Informação de Recuperação de Créditos

SIGFSNS – Sistema de Informação de Gestão do Fundo de Apoio ao Sistema de Pagamento do SNS

Intervenção Financeira do Estado

SIRIEF – Sistema de Recolha de Informação Económico-Financeira das participações do Estado.

PROGBONI -Sistema de gestão das bonificações de juros.

Gestão Patrimonial

SGD – Sistema de Gestão Documental

SGI – Sistemas de Gestão de imóveis

SIIE – Sistema de Informação dos Imóveis do Estado

Transversais

WEBTRIX – Sistema de Gestão Documental

TempoReal2000 – Sistema de controlo de assiduidade

SINGAP - Sistema Integrado para a Nova Gestão da Administração Pública

RAFE – Sistema de contabilidade e recursos humanos

BSORG e **SSD** – Sistemas de Suporte à Decisão

Homebanking – Gestão Bancaria

Análise

Os sistemas apresentados anteriormente são os blocos que vão suportar o DW, mas nem todos estão em posição de serem integrados imediatamente, sendo os sistemas transversais os mais problemáticos.

O sistema RAFE contém informação pessoal dos colaboradores, nomeadamente: vencimentos e abonos, assiduidade e classificações de serviço. A somar a estes factos, temos que o RAFE será substituído, em breve, por uma nova aplicação, o RIGORE (Rede Integrada de Gestão dos Recursos do Estado). Os sistemas associados, SSD e BSORG, serão também descontinuados com a introdução do RIGORE, pelo que não vão ser incorporados na modelação do DW.

O SINGAP é um sistema recente que ainda se encontra em fase de instalação e pertence a uma empresa privada, o que condiciona o acesso ao modelo de dados e possíveis ligações ao DW.

O *Homebanking* é a forma de acesso às contas bancarias da DGTF, o sistema está alojado no IGCP e tem um nível de segurança muito elevado e restrições ao acesso, devido ao tipo de informação que contém, pelo que não será englobado no DW.

O Sistema de Gestão Documental WebTrix regista a entrada e saída de documentos da DGTF. A natureza da informação – imagens digitais, é de pouca relevância para o DW, o que conduziu a que este sistema não fosse integrado no DW.

A área de negócio das **Regularizações e Recuperações Financeiras** tem quatro sistemas de recolha de informação. O SIGFSNS será integrado no DW. Este sistema dá suporte à informação do fundo de recuperação de pagamentos do Serviço Nacional de Saúde.

O SIRC recolhe informação sobre a recuperação de créditos no âmbito dos apoios dos Estado,

e será integrado no DW, a sua estrutura está disponível, pois foi desenvolvido na instituição.

O INVESDOC é um sistema de catalogação de documentos que é utilizado pela biblioteca. A informação contida neste sistema é de pouca relevância para as áreas de negócio da DGTF, pelo que não será integrado no DW.

O CAE é um sistema desenvolvido à medida para efectuar o controlo do Crédito Agrícola de Emergência, e será incluído no DW.

A **Intervenção Financeira do Estado** tem dois sistemas: O primeiro – SIRIEF - recolhe informação financeira das empresas participadas pelo Estado, que é colocada num servidor *MS SharePoint* e depois aglomerada em cubos OLAP, de modo a ser apresentada aos utilizadores num servidor *MS PerformancePoint*; este sistema é um DM já concluído e será integrado no DW. O segundo sistema, PROGBONI, foi desenvolvido para possibilitar o acompanhamento e gestão dos vários produtos de juro bonificado que o Estado coloca no mercado (Ex: crédito à habitação bonificado). Este sistema será integrado no DW.

A **Gestão Patrimonial** possui actualmente três sistemas, mas apenas dois se encontram em exploração. O SGD funciona somente em modo de consulta, estando em fase de concurso para ser substituído, tratando-se de um sistema de Gestão Documental e, em virtude da sua descontinuação, este não será integrado no DW.

O SGI contém informação sobre todo o património do Estado, no entanto, contém informação um pouco desactualizada e incompleta, estando a ser desenvolvidas acções de expurgo de toda a informação não conforme ou obsoleta. No intuito de recolher a informação de ocupação e propriedade foi introduzido, no inicio de 2009, o SIIE e, dando cumprimento ao Programa de Gestão e Inventariação do Património Imobiliário Público (PGPI), estes dois sistemas serão integrados no DW.

Em forma de resumo, podemos identificar na figura 4, com a cor verde, os sistemas de informação que serão integrados no DW e, a vermelho, os que vão ficar de fora.

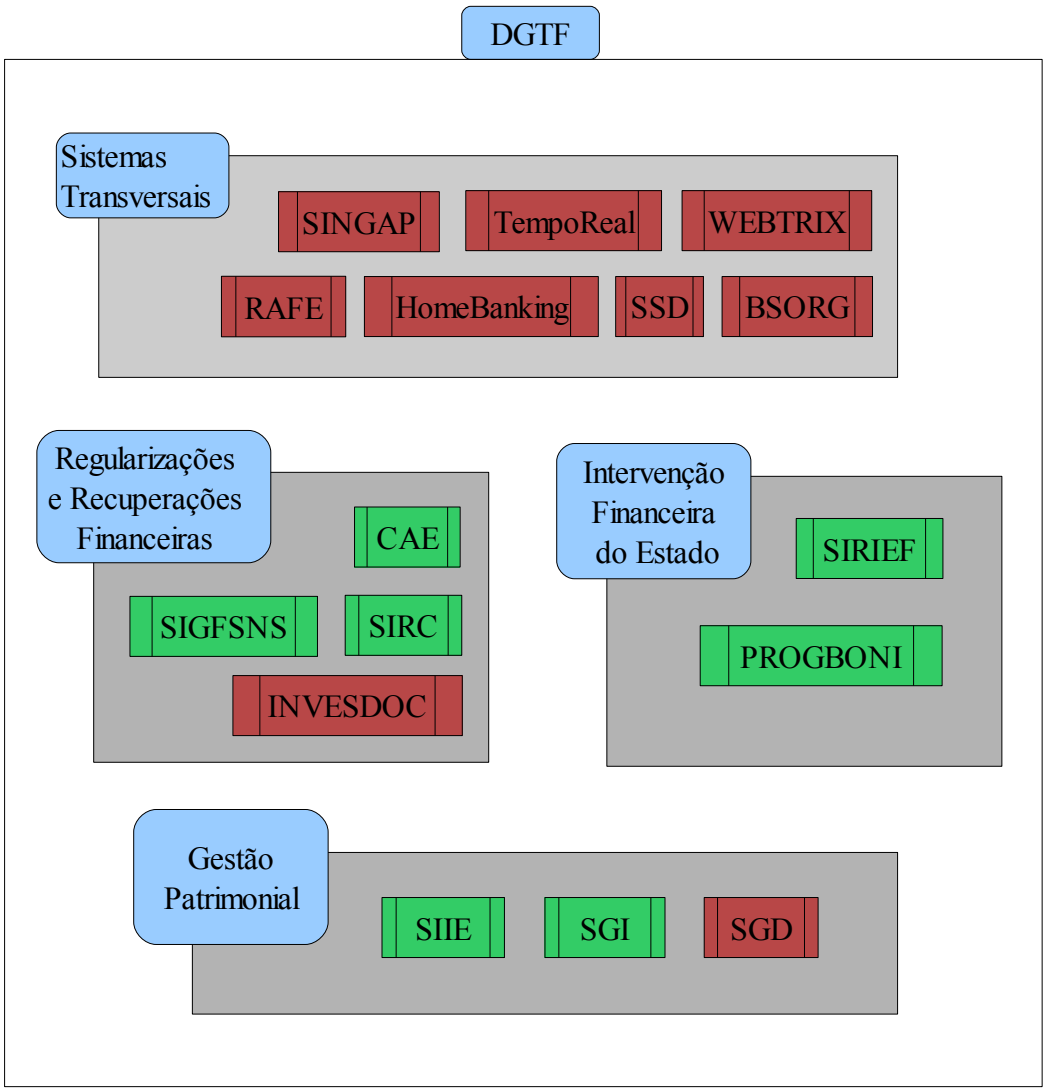

Figura 4: Sistemas de informação da DGTF

A modelação

A modelação do DW da DGTF será assente no paradigma criado por *Kimball* (2002), pelo que o DW será um conjunto de DMs que se relacionam através de dimensões *standard*, dimensões essas que, em conjunto, formam o DWB. Para modelar o DW será necessário definir o DWB.

A implementação do DW é um processo iterativo, percorrendo as várias áreas de negócio. O primeiro PN a ser incluído no DW é o da Gestão Patrimonial, o DWB apresentado será apenas correspondente a esse PN. Nas iterações seguintes serão incluídas as restantes áreas de negócio, efectuando as alterações necessárias às dimensões existentes e acrescentando novas dimensões.

5 Implementação do Data Mart

Modelação Dimensional

O processo de criação de um DM é composto por quatro passos: selecção do processo de negócio; definição da granularidade; escolha das dimensões e escolha dos factos (*Kimball*, 2002).

Selecção do processo de negócio

A escolha do primeiro processo de negócio a ser integrado no DW recaiu sobre o Programa de Gestão e Inventariação do Património Imobiliário Público (PGPI). Para dar seguimento ao disposto no Decreto-Lei n.º 280/2007, de 7 de Agosto e subsequente Resolução de Conselho de Ministros n.º 162/2008, de 24 de Outubro, que aprovou o PGPI, foi implementado o Sistema de Informação dos Imóveis do Estado (SIIE). Este sistema de informação é responsável pela recolha da informação dos imóveis e respectiva ocupação, pelos organismos da Administração Pública.

O PGPI foi escolhido por apresentar dificuldades no controlo da informação carregada e na monitorização da qualidade da mesma. A sua integração no DW terá um impacte considerável, uma vez que esta permitirá a utilização de ferramentas para a criação de relatórios automáticos e a exploração de dados pelos administradores do SIIE que monitorizam a qualidade da informação carregada.

O PGPI decorre de 2009 a 2012 e é acompanhado por um conselho coordenador composto por um representante de cada Ministério. O conselho reúne-se trimestralmente, sendo necessário criar relatórios para essas reuniões, onde seja listada informação relativa à quantidade e qualidade da informação inserida no sistema.

Identificado o processo que vai integrar o DW em primeiro lugar, é necessário efectuar um levantamento de requisitos, recolhendo as necessidades a que o DM deve responder. A informação do processo de negócio PGPI está armazenada no SIIE, que será a única fonte de dados deste DM.

O SIIE tem quatro perfis, com níveis de acesso diferentes, que podemos observar na tabela :

Perfil	Acesso
Administrador	Tem acesso a todos os dados, relatórios e menus, poder criar e alterar todos os utilizadores da aplicação, pode alterar o *layout* da aplicação;
Administrador Funcional	Tem acesso a todos os dados e relatórios, pode criar utilizadores, aprova o acesso do utilizadores externos, apaga imóveis;
Utilizador Interno	Tem acesso a todos os dados e relatórios;
Utilizador Externo	Tem acesso aos dados dos imóveis em que é proprietário ou ocupante e dos imóveis de organismos que sejam hierarquicamente dependentes, insere imóveis, proprietários e ocupantes.

Tabela 2: Nível de acesso por Perfil

Para efectivar o levantamento de requisitos foram efectuadas entrevistas aos vários utilizadores do SIIE, que foram resumidas por perfil de acesso, na tabela 3.

Perfil	Requisitos
Administrador	Informação hierarquizada sobre a utilização da aplicação. Quais os imóveis com informação em falta, errada ou incompleta agrupados por organismo?
Administrador Funcional	Relatório trimestral com taxa de introdução de imóveis e ocupantes. Informação hierarquizada do imóvel, dos proprietários e dos ocupantes. Qual o valor das rendas, por metro quadrado e por trabalhador? Qual a taxa de ocupação dos imóveis, arrendados, afectos e próprios?
Utilizador Interno	Quais os Imóveis disponíveis? Distribuição geográfica dos imóveis. Imóveis sem ocupante ou sem proprietário.
Utilizador Externo	Quais os imóveis duplicados? imóveis com informação incompleta ou errada.

Tabela 3: Requisitos por perfil de acesso

Definição da granularidade

O SIIE tem como principal objectivo a recolha de informação sobre os imóveis e sua ocupação, que é inserida pelos organismos da Administração Pública, sendo da responsabilidade de cada organismo a actualização da informação (Decreto-Lei n.º 280/2007, de 7 de Agosto).

O SIIE não é um sistema verdadeiramente transaccional; sendo sobretudo um catálogo de informação, a sua estrutura de dados não está organizada segundo a terceira forma normal, mantendo alguma informação desnormalizada de forma a ser rápido na listagem de informação.

A estrutura de dados deste sistema tem a tabela **tbl_Imovel** como ponto de partida, à qual é associada toda a informação respeitante ao imóvel, nomeadamente:

Ocupantes (**tbl_Ocupantes**) – Quem ocupa, quanta área, quantas pessoas, em que actividade;

Proprietários (**tbl_Proprietarios**) – Quem é o proprietário;

Informação do imóvel (**tbl_InfoGeral**) – Localização, Área Total, Matriz, Valor e outros.

A tabela **tbl_relacoesNPC** guarda a informação da hierarquia dos organismos da Administração Pública (AP). A tabela **tbl_NPC_NIF** guarda o tipo de entidade, sendo necessário distinguir entre privados e organismos que pertencem à AP, de entre estes, é necessário distinguir os institutos públicos ou equiparados, a administração directa e indirecta do Estado, as EPE e os organismos com autonomia patrimonial.

As tabelas **tbl_Conservatoria**, **tbl_InscricaoMatricial** e **tbl_GeoLocalizacao** guardam a informação relativa a descrição predial ao registo matricial e à localização geográfica, respectivamente.

Na Figura 5, podemos observar em pormenor a estrutura de dados composta pelas tabelas supramencionadas, bem como os campos que as constituem.

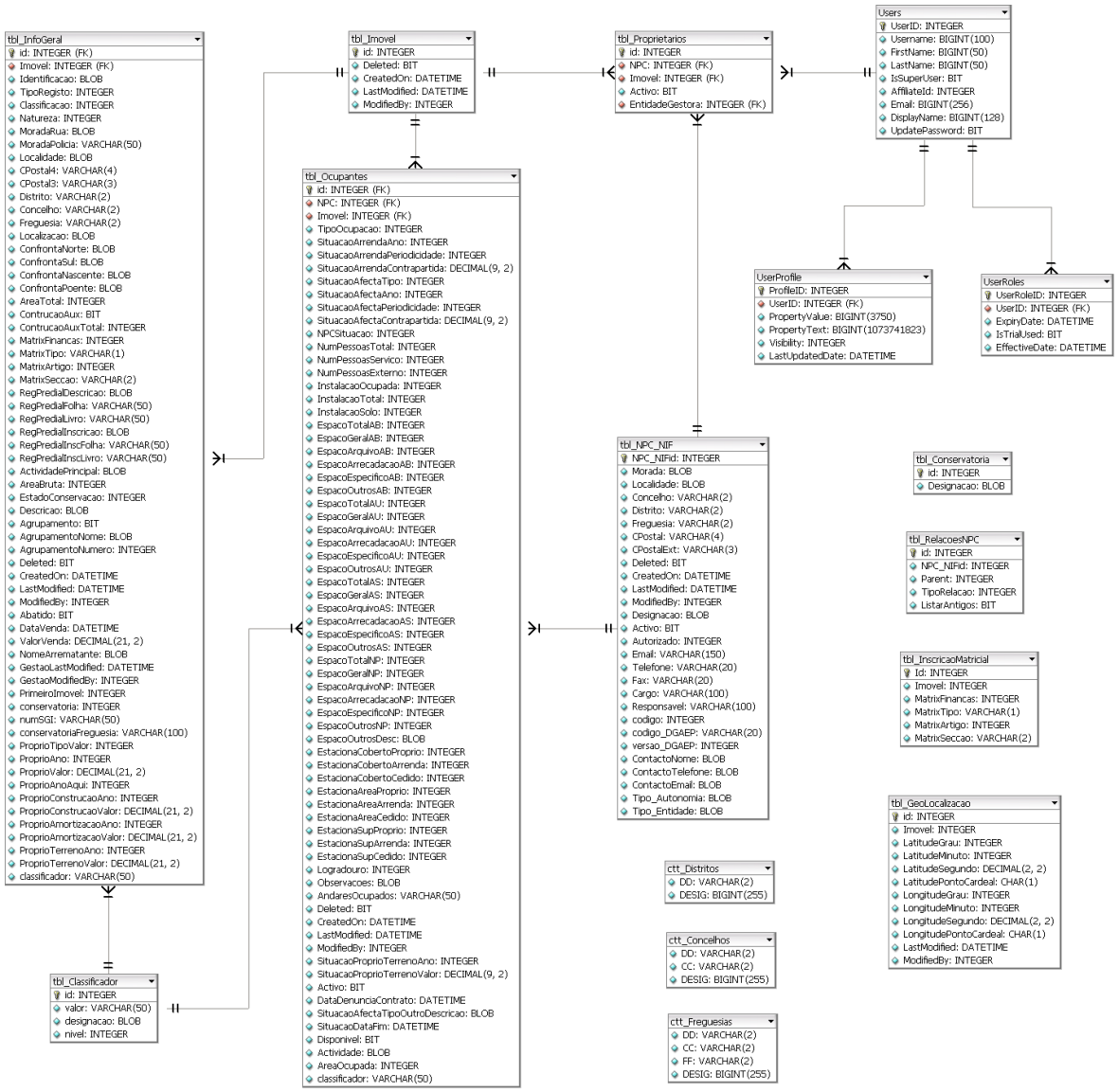

Figura 5: Modelo de dados do SIIE

O SIIE foi desenvolvido sobre um *Content Management System* (CMS) *open source* denominado *DotNetNuke* que contém bastantes funcionalidades para a gestão de sítios da Internet, como sejam a gestão de utilizadores e perfis de acesso, bem como as politicas de acesso. Deste modo a aplicação SIIE é reduzida ao essencial, tenso sido construídos módulos que foram acoplados ao *DotNetNuke*.

Na figura 5 apenas são apresentadas as tabelas relevantes para a análise em curso, tendo sido descartadas a restantes tabelas, pertencentes à plataforma *DotNetNuke*.

Numa primeira análise, e em virtude do SIIE recolher dados sobre imóveis, procurou-se definir o imóvel como grão. Contudo a necessidade de relatórios separados para o imóvel e para a sua ocupação a relação de muitos-para-muitos entre o imóvel e o ocupante e a incapacidade de utilizar o ocupante como grão para os factos do imóvel, levaram à constatação que seria mais adequado criar duas tabelas de factos, uma para o imóvel e outra para a ocupação, com granularidades diferentes.

Como observa *Kimball* (2008),

"...In the event that allocating facts down to the lowest level is impossible, the designer has no choice but to present the higher level facts in separate fact tables."

Neste sentido, foram criadas duas tabelas de factos, uma para o imóvel (F_Imovel) onde a informação guardada está ao nível do imóvel e uma tabela de factos para a ocupação (F_Ocupa) que tem o ocupante como grão.

Escolha das dimensões

A análise dos requisitos, em conjunto com a estrutura de dados do SIIE, conduziram à escolha das seguintes dimensões (Tabela 4) para a tabela de factos do imóvel e para a tabela de factos da ocupação.

Tabela de Factos	Dimensões
F_Imovel	DCF (Distrito/Concelho/Freguesia) Valor Proprietário Ocupante Ministério Classificação Data
F_Ocupacao	Ocupante Ministério Classificação Data Tipo de ocupação

Tabela 4: Dimensões das tabelas de Factos

Escolha dos factos

Os atributos considerados relevantes para as tabelas de factos foram os apresentados na Tabela 5:

Tabela de factos	Atributos
F_Imovel	Área Total Área Bruta Valor
F_Ocupa	Número Pessoas Área Bruta Área Útil Contrapartida

Tabela 5: Atributos das tabelas de Factos

Modelo Teórico

O modelo teórico do DM é apresentado na Figura 6. As tabelas de factos são representadas como rectângulos e as dimensões como elipses, sendo esta representação de alto nível da estrutura de dados o ponto de partida para a implementação física do modelo, no qual será imprescindível ter em conta as especificidade dos *softwares* utilizados para o motor de BD e para a apresentação de dados.

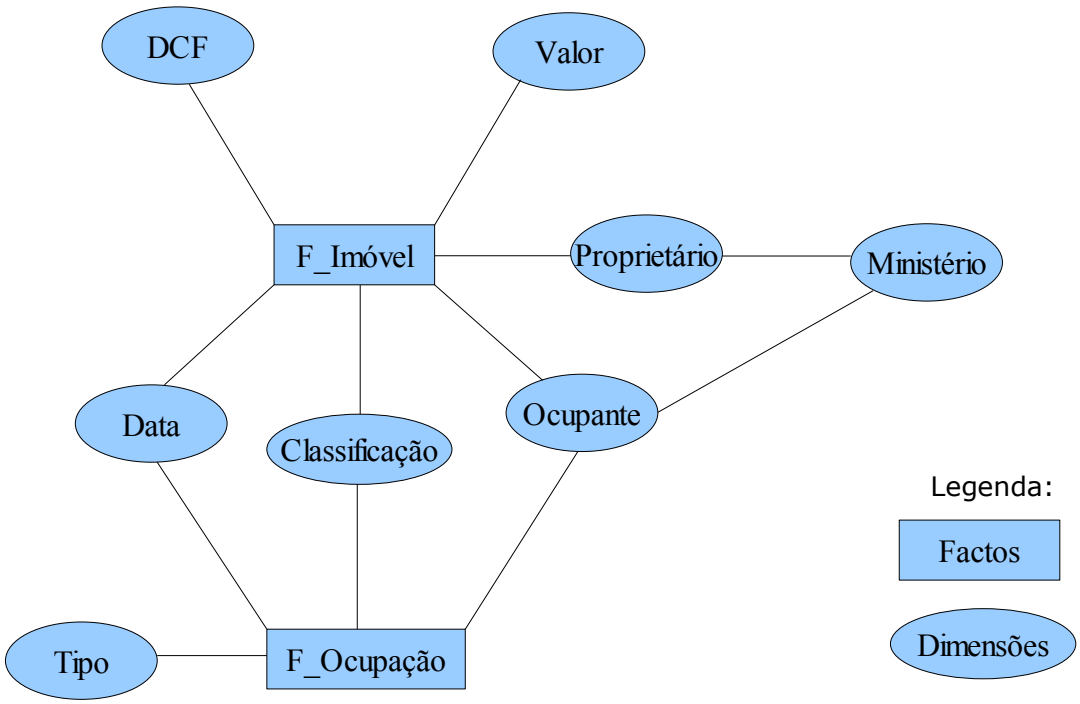

Figura 6: Modelo teórico

Modelo Físico

O modelo físico é construído a partir de uma análise detalhada das dimensões e factos do modelo teórico. Para o efeito, foi preenchido um quadro de análise para cada tabela de Dimensão (*Kimball & Corseta*, 2004), com os seguintes campos:

Campo – Nome do campo

Tipo – Indica o tipo de dados que o campo contém

Comprimento – Indica o tamanho do campo em bytes

SCD – Tipo de variação da dimensão

Fonte – Conjunto Tabela.Campo de origem no SIIE

Descrição – Breve descrição do campo

O campo que regista o *Slowly Changing Dimension* (SCD), ou seja o tipo de variação da dimensão e a importância que essa variação tem para a integridade do DW. A análise do SCD é feita campo a campo para se determinar a melhor forma de lidar com a mudança dentro de cada tabela. Podendo existir diversos níveis de SCD dentro de cada tabela. O SCD pode ser dividido em três tipos principais:

Tipo 1 - A nova informação é escrita sobre a informação antiga. Este é o método mais fácil de modelar, pois não implica nenhuma alteração às tabelas de dimensão ou factos, não há alterações de chaves ou outra informação, como desvantagem perdemos o histórico da informação.

Tipo 2 – As alterações são registadas na tabela criando um novo registo na dimensão, com uma nova chave, que é utilizada nas tabelas de factos a partir desse momento. Nas tabelas de factos são adicionados dois campos com data de início e data de termino da chave, ficando a chave actual na data de termino com a maior data suportada pelo sistema. Este procedimento viabiliza um acompanhamento da evolução ao longo de um período de tempo. Tomemos, por exemplo, a carreira de um funcionário numa empresa, em que é possível acompanhar os vários cargos e remunerações que este teve ao longo do tempo.

Tipo 3 – Este terceiro género de SCD é utilizado quando é necessário manter o valor anterior e o novo válidos. Nestes casos é acrescentada um coluna à tabela de dimensão para se guardar o valor antigo e guardar o novo no campo existente.

As tabelas do DW foram nomeadas a partir da sua designação no modelo teórico, acrescentando o prefixo "D_" para as tabeles de dimensão e "F_" para as tabelas de factos.

Para a tabela de factos do imóvel foram construídas as seguintes tabelas de dimensão:

DCF (D_DCF)

Campo	Tipo	Comp.	SCD	Fonte	Descrição
DCF_key	INT	8	-	-	Chave substituta
DCF_id	INT	6	1	ctt_Freguesias.DD ctt_Freguesias.CC ctt_Freguesias.FF	Chave original
DD_desig	VARCHAR	200	1	ctt_Distritos.DESIG	Designação do Distrito
CC_desig	VARCHAR	200	1	ctt_Concelhos.DESIG	Designação do Concelho
FF_desig	VARCHAR	200	1	ctt_Freguesias.DESIG	Designação da Freguesia

A dimensão DCF, que é implementada na tabela D_DCF, recolhe o Distrito, Concelho e Freguesia onde o imóvel fica localizado. Numa análise rápida podemos verificar que os Distritos não se alteram, sendo que os concelhos são alterados com alguma regularidade e as freguesia são alteras amiúde, no entanto, como o histórico desta informação não é relevante o SCD desta tabela será do tipo 1.

Valor (D_Valor)

Campo	Tipo	Comp.	SCD	Fonte	Descrição
Valor_key	INT	8	-	-	Chave substituta
Valor_id	INT	8	2	tbl_InfoGeral.ProprioTipoValor	Chave original
Valor_desig	VARCHAR	250	2	AT_ValorAux	Designação

A tabela de dimensão Valor contem a informação relativa ao tipo de valor que é guardado na tabela de factos do imóvel, podendo este ser: matricial, compra, avaliação ou contabilístico. Esta informação está programada directamente na aplicação; como tal, para esta poder ser

carregada no DW foi criada uma tabela auxiliar com os valores na Área de Trabalho (AT). Como é necessário manter um histórico dos valores do imóvel, esta dimensão terá um SCD do tipo 2.

Proprietário (D_Prop)

Campo	Tipo	Comp.	SCD	Fonte	Descrição
Prop_key	INT	4	-	-	Chave substituta
id	INT	4	2	tbl_Proprietarios.id	Chave original
NPC	INT	4	2	tbl_Proprietarios.NPC	Número de pessoa colectiva do proprietário
Prop_desig	VARCHAR	250	2	tbl_NPC_NIF.Designacao	Designação do proprietário
Min_key	INT	4	2	D_Ministerio.Min_key	Chave artificial do Ministério

A dimensão do proprietário guarda informação sobre o dono do imóvel que, na maioria dos casos, é o Estado, podendo também serem privados ou Institutos públicos. Esta tabela mantém também a chave para a tabela de Ministérios, com a indicação do Ministério a que o organismo pertence, no caso de ser um organismo da Administração Pública. O SCD desta tabelas será tipo 2.

Para as dimensões partilhadas entre ambas as tabelas de factos temos:

Ocupante (D_Ocupante)

Campo	Tipo	Comp.	SCD	Fonte	Descrição
Ocupante_key	INT	8	-	-	Chave substituta
id	INT	4	2	tbl_Ocupantes.id	Chave original
NPC	INT	4	2	tbl_Ocupantes.NPC	Número de pessoa colectiva do Ocupante
Ocupante_desig	VARCHAR	250	2	tbl_NPC_NIF.Designacao	Designação do Ocupante
Min_key	INT	4	2	D_Ministerio.Min_key	Chave artificial do Ministério

Na dimensão do ocupante é guardada informação sobre o organismo que ocupa o imóvel, podendo ser um privado ou um organismos da Administração Pública. Esta tabela mantém também a chave para a tabela de Ministérios, com a indicação do Ministério a que o organismo pertence, no caso de ser um organismo da Administração Pública. O SCD desta tabelas será tipo 2.

Ministério (D_Ministerio)

Campo	Tipo	Comp.	SCD	Fonte	Descrição
Min_key	INT	4	-	-	Chave substituta
id	INT	4	1	Users.UserID	Chave original
Min_desig	VARCHAR	250	1	Users.FirstName	Designação do Ministério

Esta tabela mantém a informação sobre os Ministérios existentes, podendo estes variar de Governo para Governo, sendo apenas necessário guardar informação sobre o estado actual. O registo de entidades no SIIE tem como chave o NPC, como os Ministérios não têm NPC, estes tiveram que ser criados como utilizadores da Plataforma. Os dados dos Ministérios ficam alojados na tabela Users, contrariamente aos restante organismo da AP e privados que têm a sua informação na tabela tbl_NPC_NIF.

Classificação (D_Classifica)

Campo	Tipo	Comp.	SCD	Fonte	Descrição
Classif_key	INT	4	-	-	Chave substituta
id	INT	8	1	tbl_classificador.id	Chave original
Classif_desig	VARCHAR	250	1	tbl_classificador.designacao	Designação da classificação

A Classificação de imóveis é definida na Portaria n.º 671/2000 de 17de Abril, que implementa uma listagem sistémica de classificação dos bens da Administração Publica, segundo as regras publicadas na Portaria n.º 378/94, de 16 de Junho. Para esta tabela apenas foram utilizados os classificadores de bens imóveis. Esta dimensão terá um SCD de tipo 1, pois não é necessário manter um histórico de classificações anteriores.

Data (D_Data)

Campo	Tipo	Comp.	SCD	Fonte	Descrição
DateKey	INT	4	-	-	Chave substituta
ActualDate	Data	8	1	*Script*	Data actual
Year	INT	4	1	*Script*	Ano
Quarter	INT	4	1	*Script*	Trimestre
Month	INT	4	1	*Script*	Mês
Week	INT	4	1	*Script*	Semana
DayofYear	INT	4	1	*Script*	Dia do Ano
DayofMonth	INT	4	1	*Script*	Dia do Mês
DayofWeek	INT	4	1	*Script*	Dia da Semana
IsWeekend	INT	4	1	*Script*	Fim de Semana
IsHoliday	INT	4	1	*Script*	Feriado
Comments	VARCHAR	20	1	*Script*	Comentários
CalendarWeek	INT	4	1	*Script*	Semana Civil
BusinessYearWeek	INT	4	1	*Script*	Semana Fiscal
LeapYear	INT	4	1	*Script*	Ano Bissexto

Esta é uma dimensão especial pois deve conter todas as datas possíveis. Esta tabela tem um SCD do tipo 1, pois não se prevêem alterações ao calendário. Para criar a tabela Data foi utilizado um *script* retirado do Sítio da Internet denominado *SQL Central* (2010), mas existem muito outros, com formatos diferentes, inclusive alguma aplicações de BI já trazem incorporado um gerador automático de dimensões tempo (*Microsoft*, 2010).

Para a tabela de factos do ocupante foi construída a seguinte tabela de dimensão:

Tipo (D_Tipo)

Campo	Tipo	Comp.	SCD	Fonte	Descrição
TipoOcupa_key	INT	4	-	-	Chave substituta
id	INT	8	1	tbl_Ocupantes.TipoOcupacao	Chave original
TipoOcupa_desig	VARCHAR	250	1	AT_TipoAux	Designação do tipo de ocupação

A tabela D_Tipo guarda informação sobre o tipo de ocupação que é efectuada no imóvel, esta informação pode variar num mesmo imóvel, consoante o utilização que é dada por cada um dos ocupantes. Esta informação está programada directamente na aplicação e, como tal, para poder ser carregada no DW foi criada uma tabela auxiliar com as designações na Área de Trabalho (AT).

Após a analise das diversas tabelas que compõem o DW, vamos passar à implementação propriamente dita. Para não correr riscos ao carregar dados directamente no DW, vamos usar a AT como ensaio geral para o carregamento. Foram criadas tabelas semelhantes às do DW, mas com o prefixo "AT_", o carregamento será efectuado primeiro nesta tabelas e só depois as tabelas do DW serão carregadas a partir da AT.

Na AT foram criadas duas tabelas auxiliares para carregamento das designações do tipo de ocupação dos imóveis e para a designação do tipo de valor designadas AT_TipoAux e AT_ValorAux, respectivamente.

O SIIE tem nas sua tabelas vários campos de auditoria, nomeadamente: *LastModified, Modifiedby e CreatedOn*. Este campos serão utilizados para controlar os carregamentos no DW. Para o efeito foi criada uma tabela denominada AT_Quorum que serve para manter a data do último carregamento, sendo comparada com o campo *LastModified* de cada registo para saber-mos quais os registos que foram alterados desde o último carregamento.

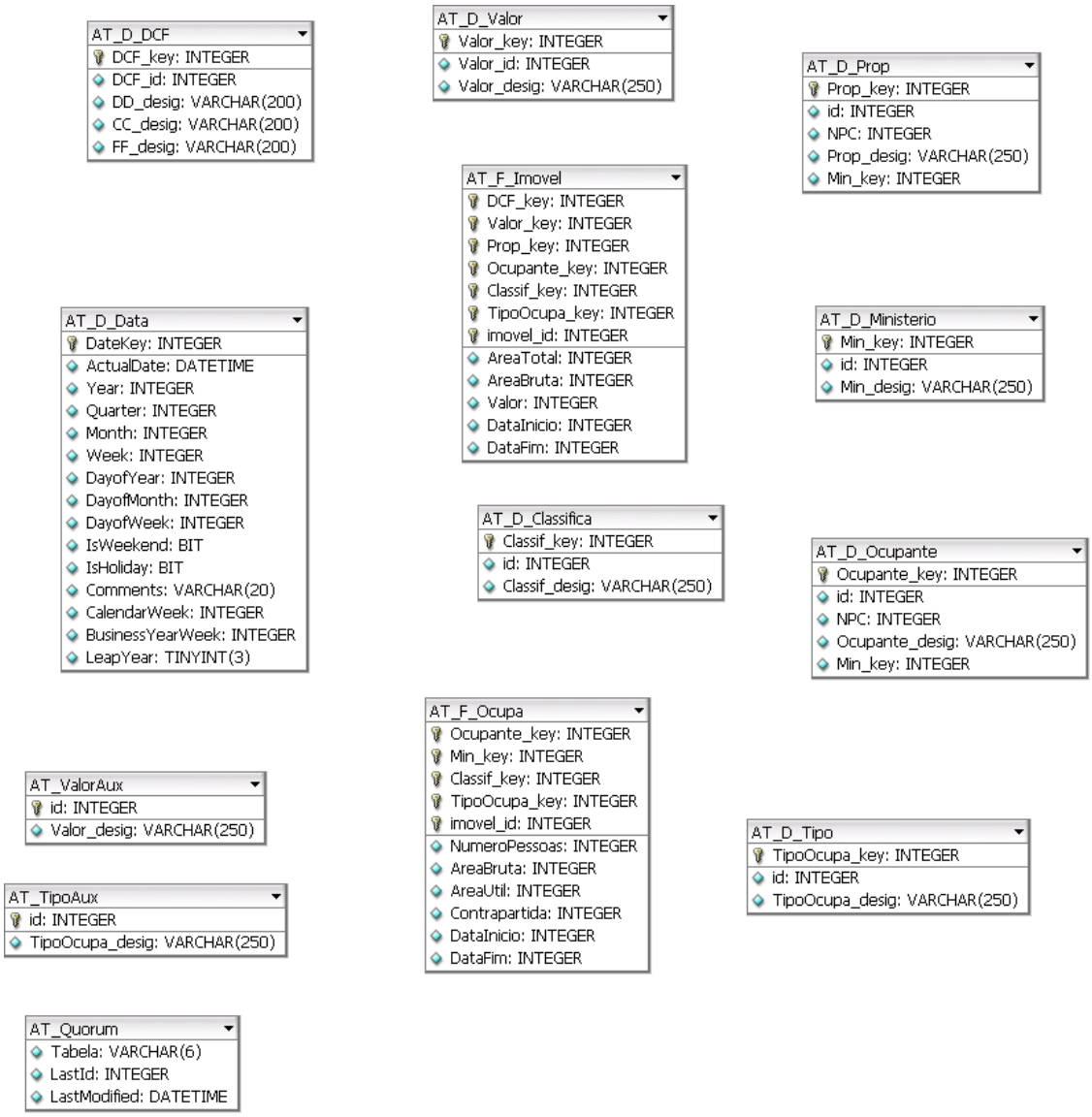

Figura 7: Modelo físico da AT

Na figura 7 podemos observar o modelo da AT. As tabelas ainda não têm os ligações relacionais porque são carregadas parcialmente, com os dados novos, e isso pode causar problemas com as chaves primárias de cada tabela.

Definição do DWB

O DWB será definido primeiramente a um nível mais elevado, através de uma matriz, onde se cruzam as dimensões, na vertical, com os componentes do negócio, na horizontal.

A Tabela 6 apresenta a matriz do DWB, após a modelação do primeiro DM.

		Dimensões							
		DCF	Valor	Proprietário	Ministério	Ocupante	Classificação	Data	Tipo Ocupação
Factos	Imóvel	X	X	X	X	X	X	X	
	Ocupação				X	X	X	X	X

Tabela 6: Matriz do Data Warehouse Bus

Implementação Física

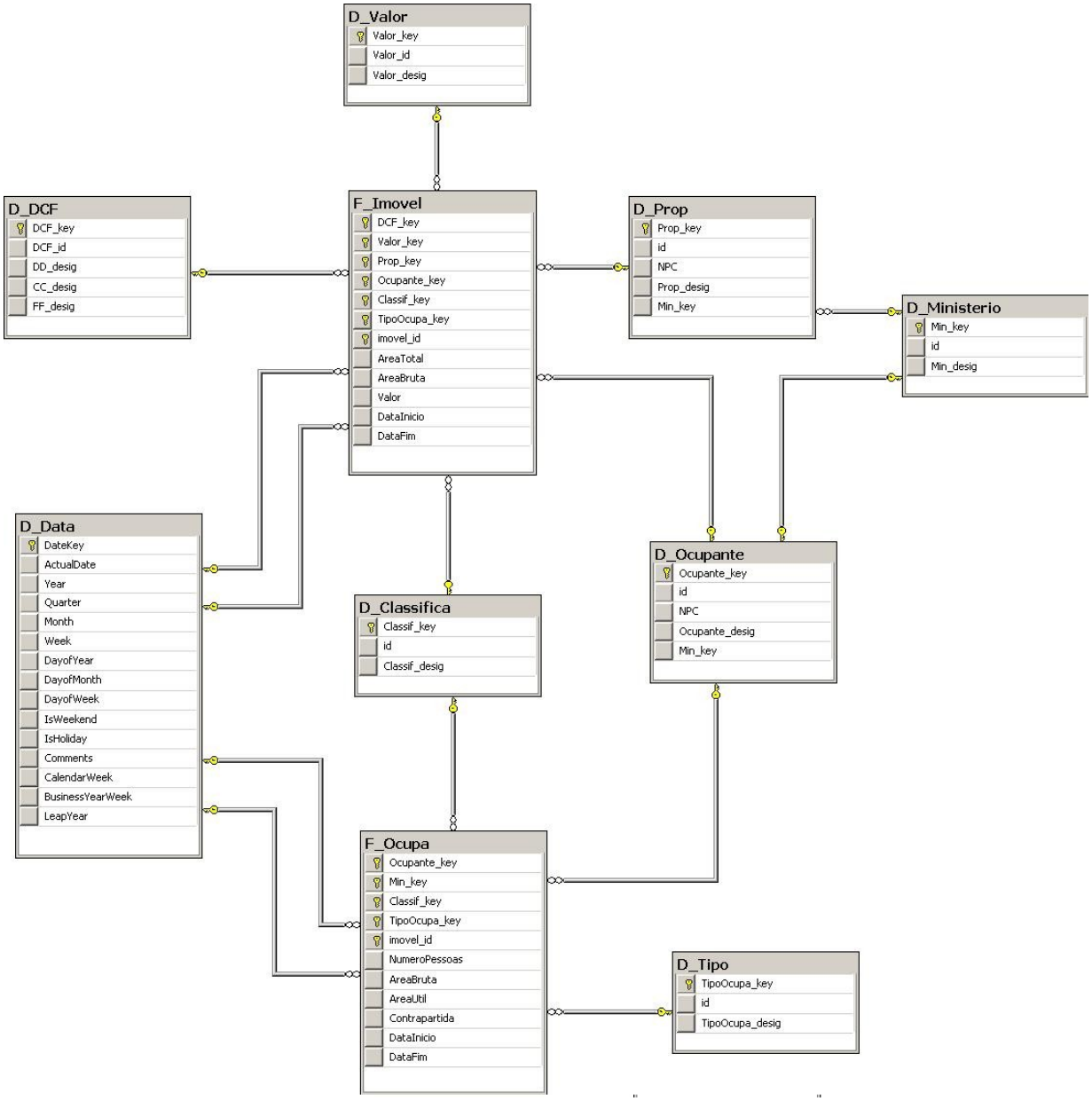

Figura 8: Diagrama do DM em *SQL Server* 2005

Na figura 8 temos o diagrama do primeiro DM implementado em *SQL Server* 2005. Com esta definição terminamos a primeira iteração da modelação do DW. Antes de partir para a integração da segunda área de negócio temos que carregar este DM com dados.

Carregar o Data Mart

No seguimento do ponto anterior, onde foi construído o DM, será necessário carregar a informação existente, no sistema de origem, no DM. Esta fase é designada de ETL proveniente da designação anglo-saxónica *Extraction, Transformation and Loading*, em Português, Extracção, Transformação e Carregamento. Como o nome indica está fase está dividida em três partes.

1) Extracção – A informação é recolhida dos sistema de origem e depositada numa área de trabalho (AT), em Inglês, denominada *Staging Area* (SA) onde fica disponível para o passo seguinte.

2) Transformação – Na segunda parte, os dados que estão na SA vão ser transformados de modo a ficarem compatíveis com a estrutura do DM, limpando os dados, verificando a integridade, fazendo agregações de informação e validando o resultado obtido.

3) Carregamento – O último passo consiste em exportar a informação anteriormente preparada para o DM.

O processo de ETL não é rígido, podendo por vezes unir ou saltar fases. Para efectuar o ETL do projecto foi escolhido o *software Pentaho Data Integration* (PDI), esta aplicação tem funcionalidades avançadas, permitindo num só passo executar as três fases do ETL (*Pentaho*, 2010).

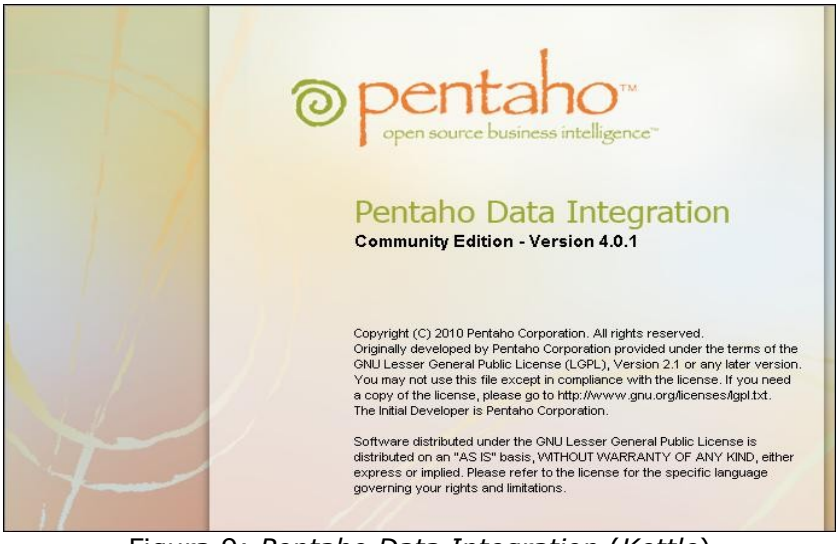

Figura 9: *Pentaho Data Integration* (*Kettle*)

O PDI organiza os diversos passos necessários para cumprir um processo ETL em *Jobs*, estes podem conter outros *Jobs* ou *Transformations*, que por sua vez são constituídos por *Steps*, a unidade atómica. Os *Steps* estão divididos em duas grandes áreas, uma para os *Jobs* onde encontramos *Steps* com funções de mais alto nível, para enviar correio electrotécnico, gerir ficheiros e pastas ou despoletar *Transformations*, como podemos ver na Figura 9.

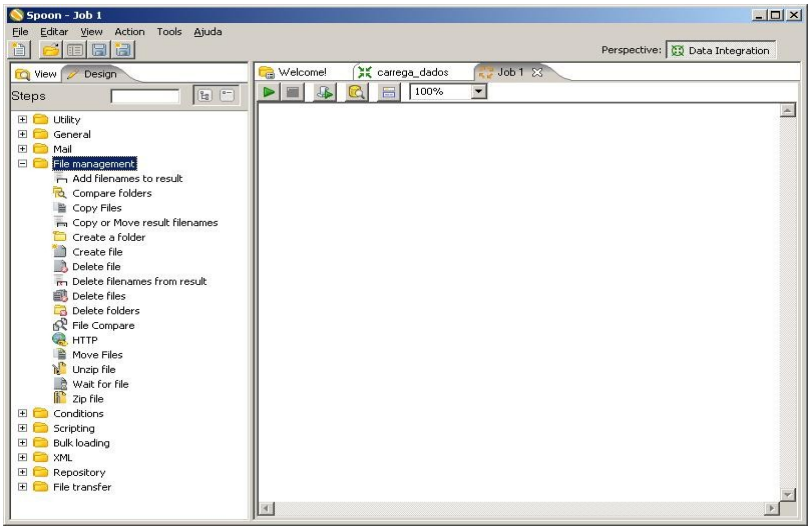

Figura 10: *Steps* para *Jobs*

O segundo bloco de Steps é destinado a funções de baixo nível, para construção das *Transformations*, visíveis na figura 10, onde podemos encontrar funções para carregar/descarregar dados de uma grande variedade de fontes, conversão de dados, selecção baseada em regras, entre outras.

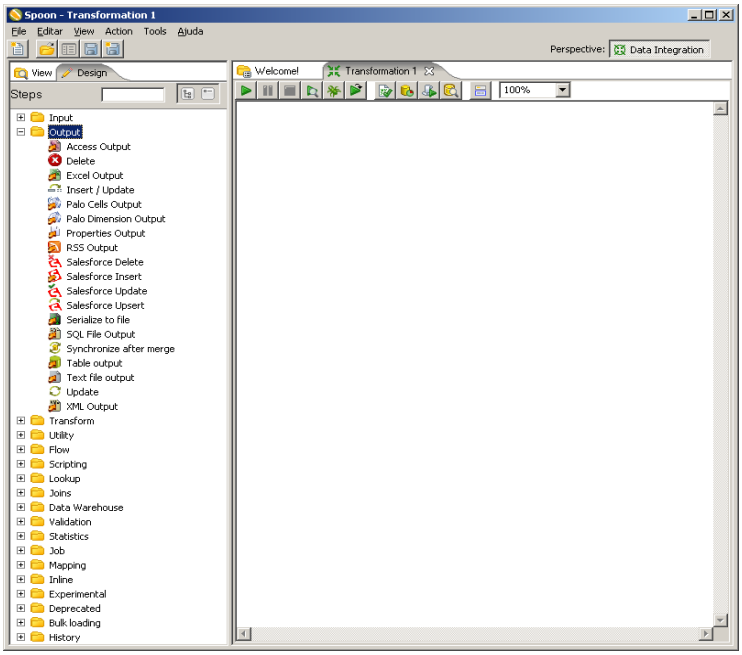

Figura 11: *Steps* para *Transformations*

O processo de ETL será programado para correr automaticamente todas as noites, no entanto o ETL do SIIE tem algumas particularidades que passamos a descrever:

A informação da dimensão tempo é construída a partir de um *script*, estando a actual tabela preparada com datas até 2020. Qualquer novo carregamento será feito a partir do *script*, não havendo lugar a ETL automático.

Os valores da dimensão DCF são importados a partir de ficheiros que os CTT publicam semestralmente, assim, a tabelas D_DCF terá o seu próprio *Job* que será corrido manualmente cada vez que for carregado um ficheiro novo no SIIE.

As dimensões Valor e Tipo estão programadas na aplicação, pelo que a sua alteração necessita da intervenção de um elemento da equipa de desenvolvimento. A sua actualização não pode ser efectuada de forma automática sendo criados *Jobs* próprios para cada uma, que devem ser corridos manualmente.

A classificação dos imóveis depende de alterações legais à Portaria n.º 671/2000 de 17de Abril, como tal este processo fica de fora do automatismo, sendo criado um *Job* próprio que será corrido manualmente.

Os Ministérios são alterados de Governo para Governo ou, em caso de remodelações que impliquem alterações à estrutura ministerial, esta dimensão terá o seu próprio *Job* manual de carregamento.

As restantes dimensões (Proprietario, Ocupante) e as tabelas de factos serão automatizadas e preparadas para correr todos os dias.

ETL automático

O processo de ETL automático foi construído a partir de um *Job* principal denominado ETL_SIIE e, dentro deste, foram colocados quatro outros *jobs*, um por cada tabela de dimensão e factos. Os primeiros a correr serão as dimensões e ,depois, os factos.

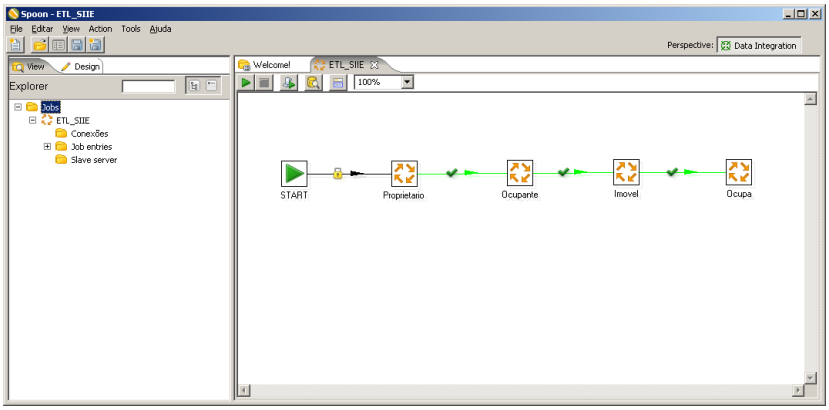

Figura 12: ETL_SIIE

O *Job* Proprietário é composto por duas *Transformations* e um *script* de SQL, como podemos ver na figura 13.

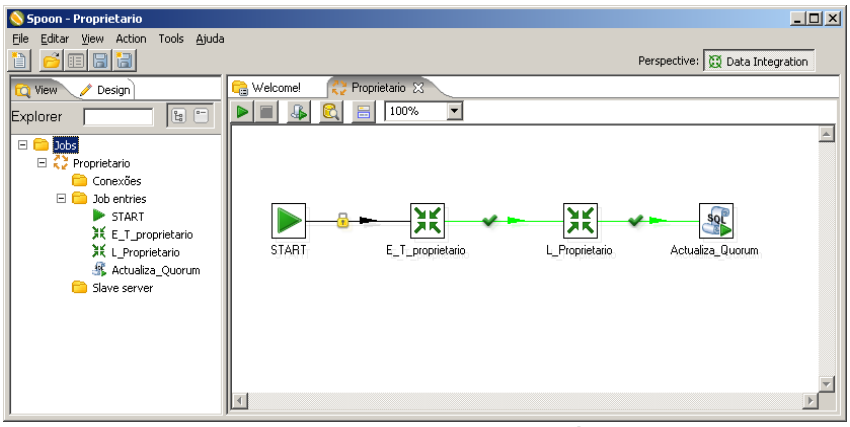

Figura 13: O *Job* Proprietário

Para nomear as *Transformations* foram usados prefixos consoante a processo que efectuavam. Na figura 13, podemos ver a *Transformation* E_T_Proprietario, em que o prefixo "E_T_" indica que se trata de um processo de Extracção e Transformação, no caso da *Transformation* L_ Proprietario trata-se de um carregamento (*Loading*).

Na figura 14, temos o conteúdo da *Transformation* E_T_Proprietario, que recolhe informação em várias tabelas e a coloca na AT.

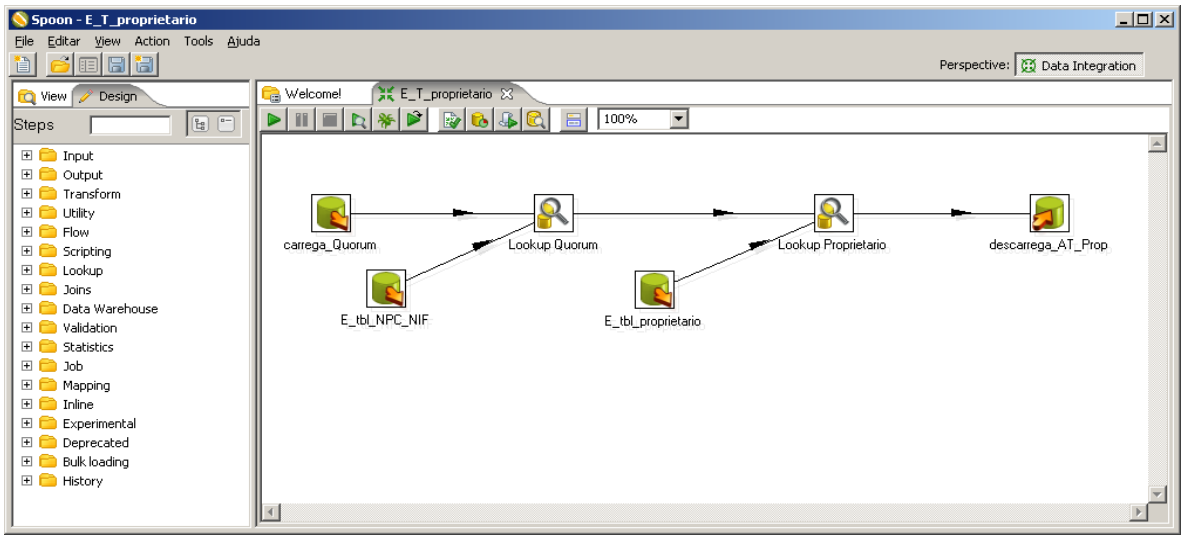

Figura 14: *Transformation* E_T_Proprietario

Temos então, um carregamento da informação da tabela AT_Quorum, que podemos analisar em mais pormenor na figura 15.

O *Step* carrega_Quorum tem dois elementos principais, a ligação à BD e a instrução em SQL que vai ser executada.

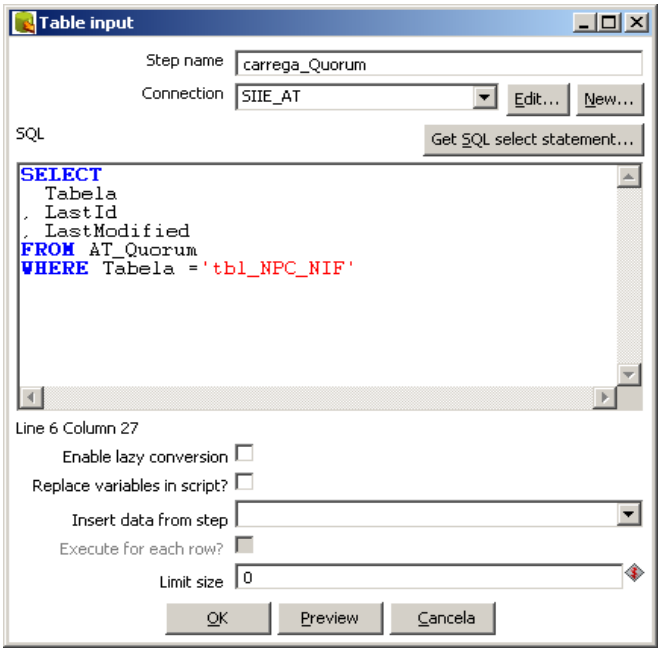

Figura 15: carrega AT_Quorum

Na figura 16 pode-mos ver o *Step* LookUp_Quorum, em que é efectuada uma comparação entre os campos *LastModified* e é efectuado um carregamento dos dados que respeitem esse condição

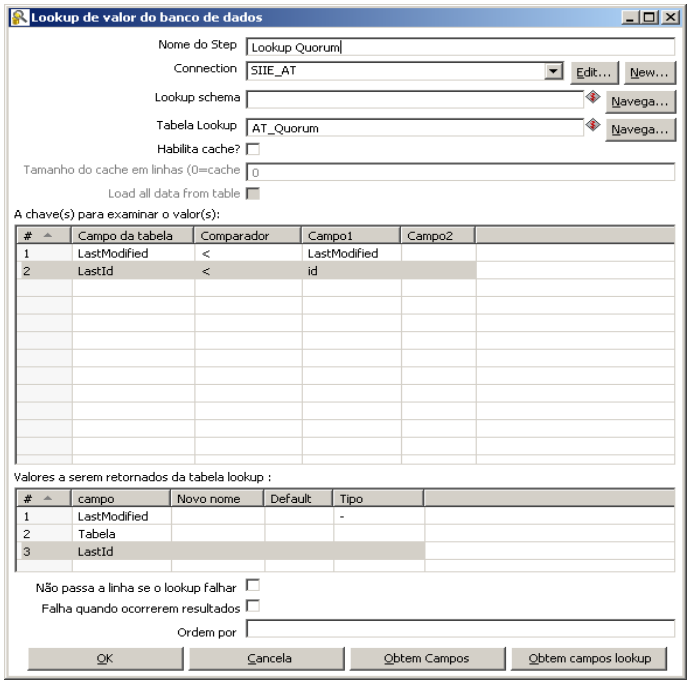

Figura 16: LookUp_Quorum

A figura 17 representa o *Step* LooUp_Proprietario onde é efectuado uma comparação e é carregada informação relativamente à designação do Proprietário.

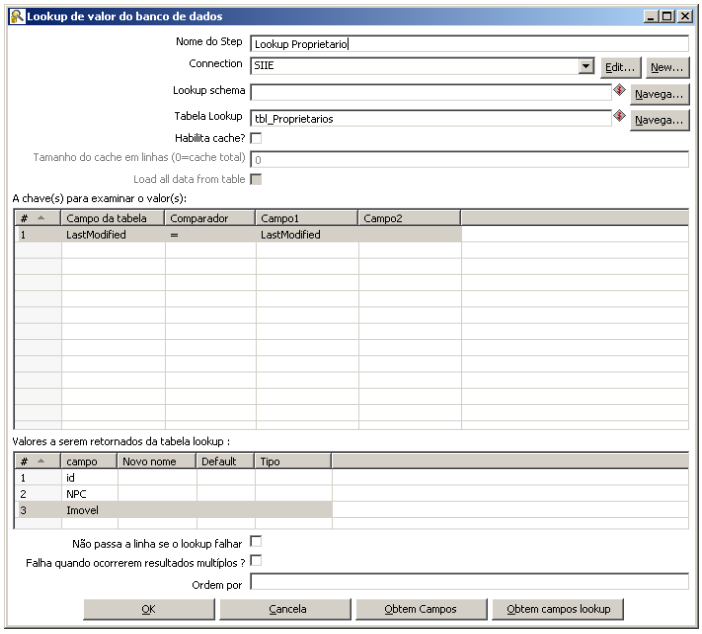

Figura 17: LooUp_Proprietario

Depois de recolhida a informação pretendida esta é descarregada na AT, para validação e posterior carregamento no DM.

Este passo é executado com a ajuda de um *Step Table output* que recebe informação do step anterior e a deposita na tabela AT_D_Prop.

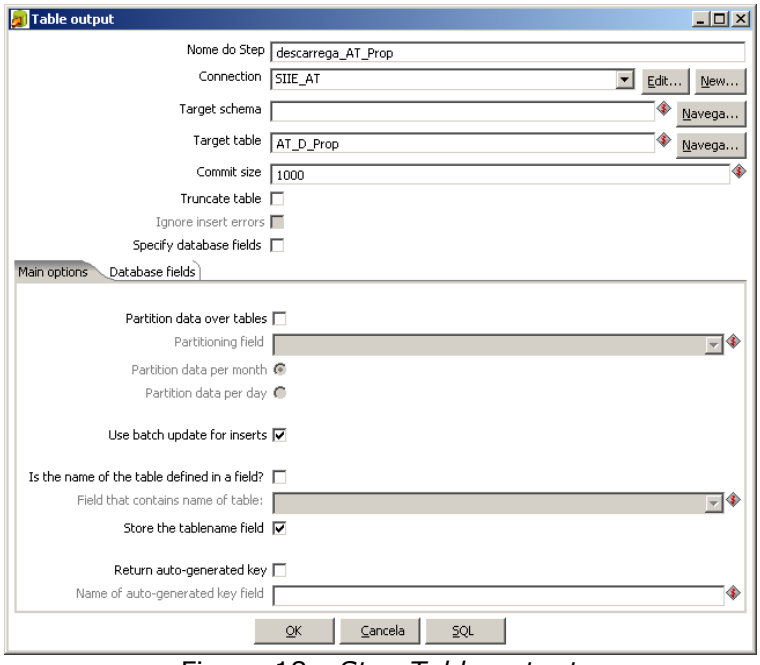

Figura 18: *Step Table output*

A *Transformation* seguinte è de carregamento, L_Proprietario, sendo bastante simples, na

medida em que apenas lê dados da AT e descarrega-os no DM.

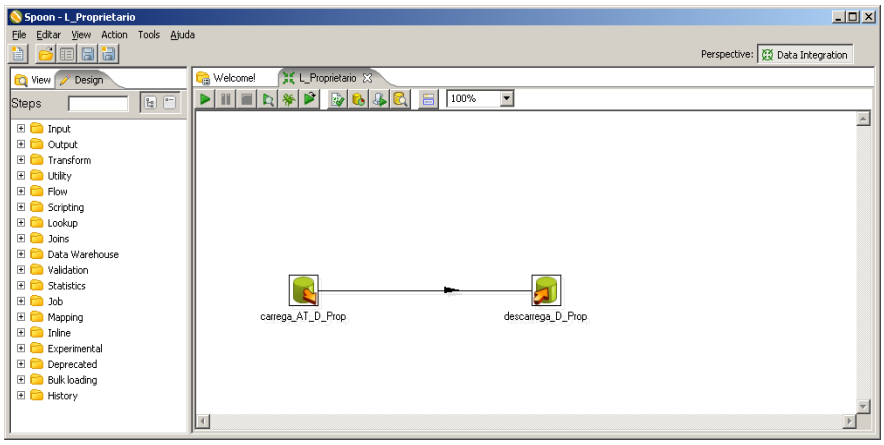

Figura 19: Carregamento de dados no Data Mart

O *Job* de ETL da dimensão D_Prop só termina quando é efectuado um script SQL que actualiza a tabela AT_Quorum com os novos valores de Id e *LastModified*.

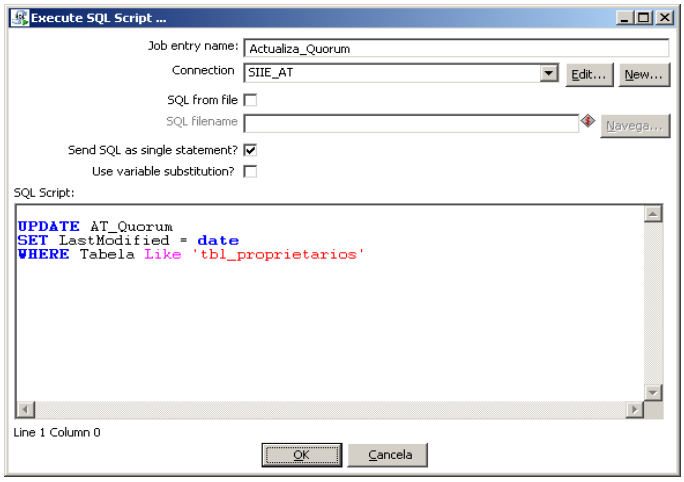

Figura 20: Actualização da tabela AT_Quorum

Para evitar repetições, não foram analisados todos os *Jobs*, apenas os que apresentam características diferenciadoras foram descritos.

6 Resultados

Do Data Warehouse

A aplicação da metodologia escolhida à modelação do DW foi concluída com sucesso, tendo sido possível definir a primeira iteração do DWB. Na Tabela 7, podemos observar a matriz de alto nível do *Bus,* onde são cruzadas as componentes do PN com as Dimensões.

		Dimensões							
		DCF	Valor	Proprietário	Ministério	Ocupante	Classificação	Data	Tipo Ocupação
Factos	Imóvel	X	X	X	X	X	X	X	
	Ocupação				X	X	X	X	X

Tabela 7: Matriz do DWB

O resultado final da modelação só foi visível depois de implementado o primeiro DM, sendo assim possível preencher os atributos de cada Dimensão. Na figura 22 observamos a primeira iteração física do DWB.

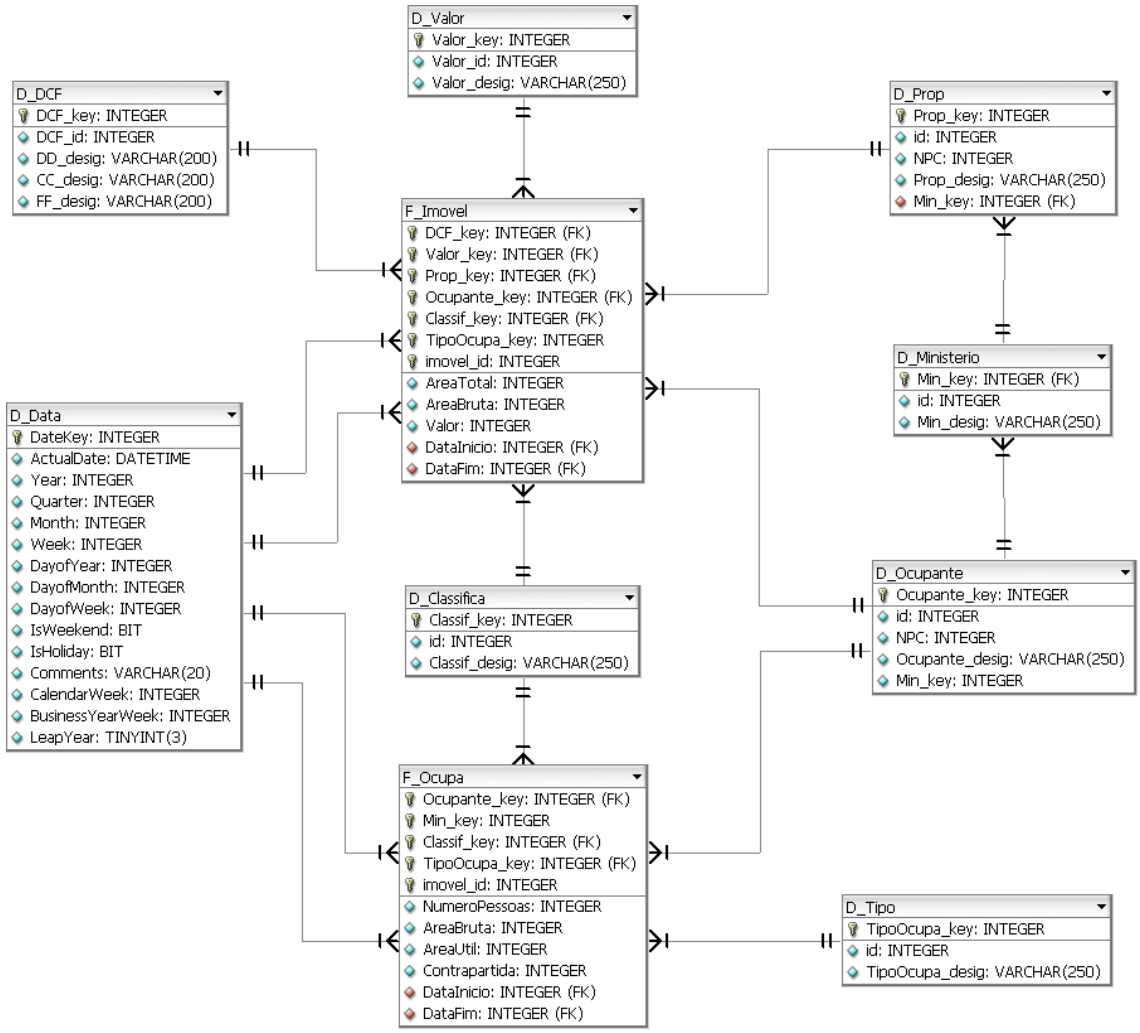

Figura 21: A primeira iteração do DWB

Do Data Mart

A implementação do DM conclui-se com a primeira iteração da metodologia de desenvolvimento que foi escolhida. Após a definição do modelo físico do DM, este foi carregado com dados do sistema de origem (SIIE) terminando, desta forma, a implementação.

Na figura 23 podemos observar o modelo físico do DM.

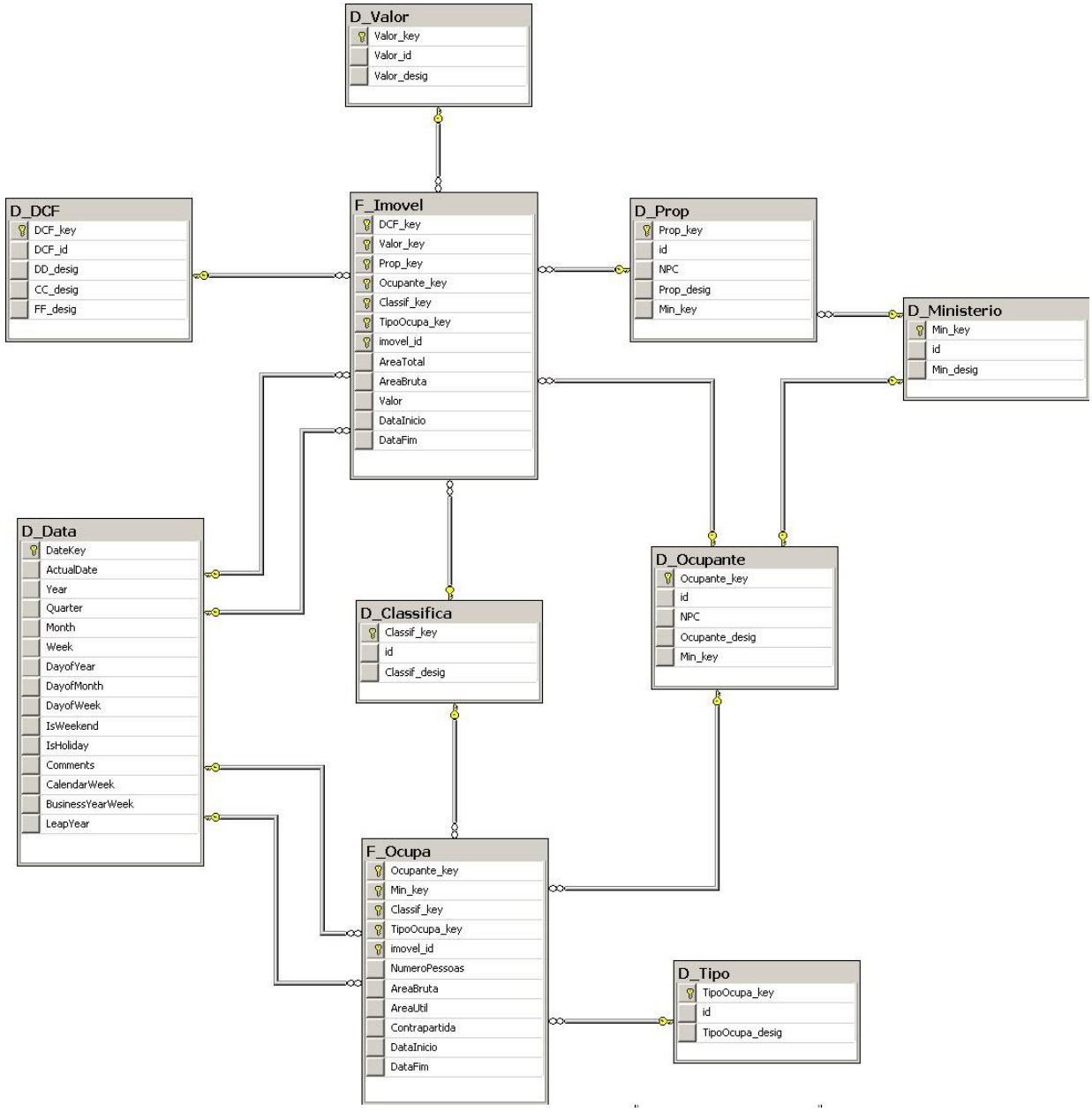

Figura 22: Modelo físico do DM

Relatórios

A capacidade de gerar relatórios de forma independente e autónoma do Serviço de Informática representa uma mais-valia. A plataforma *Pentaho* tem uma aplicação própria para a criação de relatórios, o *Pentaho Report Designer*, que disponibiliza funcionalidades equivalentes a *softwares* comerciais, nomeadamente: um assistente para a criação de relatórios, que acompanha o utilizador através dos vários passos necessários à criação de um relatório; modelos de relatórios, para que estes possam ser utilizados por todos os utilizadores, com uma identidade visual uniforme; relatórios interactivos, respondendo a parâmetros escolhidos pelo utilizador final; repositório central de relatórios, que permite aos utilizadores a partilha e o acesso imediato aos relatórios. O *Pentaho Report Designer* tem um interface gráfico do tipo *drag-and-drop* que permite um fácil manuseamento dos campos de informação, ligações a BD e, funcionalidades de cálculo, entre outras.

O suporte de ligação às fontes de dados mais comuns vai permitir a utilização dos conhecimentos adquiridos nesta plataforma noutros sistemas, uma vez que será possível ligar o *Pentaho Report Designer* aos repositórios de outras aplicações, internas ou externas à DGTF.

7 Conclusão

Discussão

O conceito de *Data Warehouse* foi bem recebido pelos responsáveis da Gestão Patrimonial da DGTF, dado que a sua concretização preencheu uma área que não estava bem explorada: a gestão do processo de negócio. Tornou-se assim evidente que, através de uma monitorização constante do processo é possível intervir numa fase inicial dos problemas, lançando medidas correctivas que evitem problemas maiores.

O SIIE tem um problema grave de qualidade da informação, visto que os imóveis do Estado têm a sua informação física dispersa por vários organismos, sendo que algum desse património não está devidamente registado nas conservatórias, ou sofreu uma série de alterações, após o registo, que não foram actualizadas. Deste modo, torna-se impossível encontrar uma chave primária para a entidade imóvel. Este problema causa imensos imóveis repetidos, sendo necessário alertar os organismos ocupantes para efectuarem as devidas correcções. O DW permite que sejam executados relatórios automáticos para detectar situações incorrectas ou incoerência de dados.

O sistemas de relatórios também possibilita a execução de relatórios *ad-hoc*. Estes são de grande utilidade para responder às várias solicitações que são efectuadas pela tutela, no caso, o Secretário de Estado do Tesouro e Finanças, ou por entidade externas, como seja o caso do Tribunal de Contas ou Grupos Parlamentares da Assembleia da República.

Os relatórios produzidos pelo sistema não são apresentados porque contém informação sigilosa. A autorização que foi concedida apenas contemplava a utilização dos dados para efeitos de modelação, não podedo ser apresentados valores, quer na foma isolada ou agregada.

Não obstante, a exploração da informação e a facilidade com que esta pode ser manuseada tem atraído mais utilizadores a este projecto, fazendo prever uma rápida adesão dos restantes PN.

A utilização da metodologia *Kimball,* em oposição à *Inmon,* revelou ser a mais acertada, pois facilitou a apresentação rápida de resultados, associada à flexibilidade do DWB que permite a introdução de novos processos. Assim, será possível continuar o desenvolvimento, sem interromper o funcionamento do DW actual.

Um problema que foi levantado na fase final do projecto e que não está contemplado, foi a formação dos utilizadores finais; embora as ferramentas apresentadas sejam intuitivas e com interfaces apelativos, existe sempre uma curva de aprendizagem que poderia ser atenuada

com acções de formação, para facilitar os primeiros passos e explorar as opções disponíveis com recurso a casos práticos.

Considerações finais

Os dois objectivos apresentados no início deste projecto foram cumpridos; tanto da perspectiva teórica, onde foi modelado o DW e o respectivo DWB, que dá suporte às dimensões *standard*; como do ponto de vista prático, onde foi implementado um DM para o PGPI.

A construção de um DW é um processo iterativo, razão pela qual nunca está verdadeiramente terminado, sendo um processo contínuo de aperfeiçoamento da estrutura de dados. A natureza dinâmica de alguns sistemas ou a sua implementação por fases, como é o caso do SIIE, provoca alterações regulares no DW, sendo necessário adaptar o DWB a novas dimensões ou, em casos extremos, alterar dimensões. Para o futuro, fica a construção dos DMs dos restantes processos de negócio até estarem cobertos todos os processos de negócio da DGTF.

O facto de este projecto ter apenas uma pessoa a trabalhar em *part-time* levou a que não fosse possível ir muito além do que estava previsto nos objectivos. Contudo existe a percepção de que o DW pode ser melhor rentabilizado se forem acopladas outras ferramentas, que seguidamente passamos a enumerar:

A implementação de um servidor de cubos OLAP. A agregação da informação num, ou vários, cubos OLAP permite aos utilizadores uma aumento na versatilidade da manipulação da informação, utilizando tabelas dinâmicas na folha de cálculo *Excel* ou via o ambiente web disponibilizado na plataforma *Pentaho*, denominado *Pentaho Analysis*. O servidor de cubos também diminui o tempo dos cálculos, pois a informação já esta agregada. A construção de relatórios com informação proveniente de um cubo OLAP permite utilizar as funcionalidades de *Slice*, *Dice*, *Drill-up* e *Drill-Down*, funcionalidades essas que permitem "navegar" até ao pormenor da informação atómica.

O servidor de informação geográfica. A plataforma *Pentaho* disponibiliza um interface de integração com o *Google Maps*, o que permite georeferenciar a informação a partir da morada, podendo esta ser recolhida no DW. A representação da informação em formato geográfico facilitará o processo de eliminação de imóveis duplicados.

As novas ferramentas disponibilizadas no âmbito do DW abrem vários caminhos de investigação, no campo da análise de dados e da Informação Geográfica. A aplicação de metodologias como o *Balanced Score Card* ou o *6-SIGMA* aos processos da DGTF seria uma mais-valia.

Referências

Livros

Adelman S. e Moss L. T. (2000). "*Data Warehouse Project Management*", I. Technology, Addison-Wesley

Boar B. H. (1997). "*Building, Using, and Managing the Data Warehouse-Understanding Data Warehousing Strategically*", Prentice Hall PTR

Boehm, B.W. (1988). "*A Spiral Model of Software Development and Enhancement*", Computer, Volume 21, Issue 5, 61-72.

Chaudhuri S. e Dayal U. (1996). "*An Overview of Data Warehousing and OLAP Technology*", VLDB Conference

Breslin M. (2004). "*Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models"*. Recuperado em 21 de Outubro de 2010, de http://www.tdwi.org/research/display.aspx?ID=6991

Greenfield L. (2002) ."*The Case for Data Warehousing*". Recuperado em 30 de Outubro de 2010, de http://www.dwinfocenter.org/casefor.html

Gupta, V. R. (1997). "*An Introduction to Data Warehousing*", Ed., 1997

Hammer J., Garcia-Molina H., Jennifer W., Labio W. e Zhuge Y. (1995). "*The Stanford Data Warehousing Project*", IEEE Data Engineering Bulletin, Volume 18 , 41-48

INE – Instituto Nacional de Estatística (2006). "*RIAP - Perguntas Frequentes"*. Recuperado em 30 de Outubro de 2010, de http://webinq.ine.pt/public/files/inqueritos/riap/perguntasfrequentes.aspx?Id=188

Inmon, W.H. (1994). "*Using the Data Warehouse*", John Wiley & Sons, Inc.

Inmon, W.H. (1997). "*Managing the Data Warehouse*", John Wiley & Sons, Inc.

Inmon, W.H. (2000)a. "*Data Mart Does Not Equal Data Warehouse*". Recuperado em 30 de Outubro de 2010, de http://www.information-management.com/infodirect/19991120/1675-1.html

Inmon, W.H. (2000)b. "*A Data Warehouse Development Methodology*". Recuperado em 30 de Outubro de 2010, de http://www.inmoncif.com/view/17

Inmon, W.H. (2000)c. "*The Future of Data Warehousing: Alternative Storage*". Recuperado em 30 de Outubro de 2010, de http://www.inmoncif.com/view/14

Inmon, W.H. (2002). *"Building the Data Warehouse, 3th ed."*. John Wiley and Sons, Inc.

Inmon, W. H. , Imhoff, C. e Sousa, R. (1998). "*The Corporate Information Factory*". Recuperado em 30 de Outubro de 2010, de http://www.inmoncif.com/library/cif/

Kimball R., Reeves L., Ross M. e Thornthwaite W. (1998). "*The Data Warehouse Lifecycle Toolkit - Expert Methods for Designing, Developing, and Deploying Data Warehouses".* John Wiley & Sons, Inc.

Kimball, R. e Ross, M. (2002). *"The Data Warehouse Toolkit : the complete guide to dimensional modeling, 2nd ed."*. John Wiley and Sons, Inc.

Kimball, R. e Corseta, J. (2004). *"The Data Warehouse ETL Toolkit : practical techniques for extracting, cleaning, conforming, and delivering data"*. Indianápolis: Wiley Publishing, Inc.

Kimball R., Ross M., Thornthwaite W., Mundy J. e Becker B. (2008). *"The Data Warehouse Lifecycle Toolkit, 2nd ed."*. Indianápolis: Wiley Publishing, Inc.

Lee M. L., Lu H., Ling T. W. e Ko Y. T. (1999). "*Cleansing Data for Mining and Warehousing*", 10th International Conference on Database and Expert Systems Applications, Florence, 751-760

Maletic J. I. e Marcus A. (2000). "*Data Cleansing: Beyond Integrity Analysis*", Conference on Information Quality, Boston, 200-209

Microsoft (2010). "Creating a Time Dimension by Generating a Time Table". Recuperado em 30 de Outubro de 2010, de http://msdn.microsoft.com/en-us/library/ms174832.aspx

Poe V. ,Klauer P. e Brobst S. (1998). "*Building a Data Warehouse for Decision Support*", 2nd Ed., Prentice Hall

Pentaho (2010). "*Pentaho Community Home"*. Recuperado em 30 de Outubro de 2010, de http://community.pentaho.com/

Porter, M. E. e Millar, V. E. (1985). "How Information Gives You Competitive Advantage", Harward Business Review

Shim J. P., Warkentin M., Courtney J. F., Power D. J., Sharda R. e Carlsson C. (2002). "*Past, present, and future of decision support technology*", Decision Support Systems, Volume 33, 111-126

SQL Central (2010). "*Create and Populate Time Dimension*". Recuperado em 30 de Outubro de 2010, de http://www.sqlservercentral.com/scripts/Data+Warehousing/30087/

Subramanian A., Smith L. D. e Nelson A. C. (1996). "*Strategic Planning for Data Warehousing in the Public Sector*", 29th Annual Hawaii International Conference on System Sciences, Hawaii, 54-60

Legislação

Decreto n.º 22, de 16 de Maio de 1833

Decreto-Lei n.º 49-B/76, de 20 de Janeiro

Decreto–Lei n.º 158/96, de 3 de Setembro

Decreto-Lei n.º 205/2006, de 27 de Outubro

Decreto-Lei n.º 280/2007, de 7 de Agosto

Decreto Regulamentar n.º 21/2007, de 29 de Março

Resolução de Conselho de Ministros n.º 40/2004, de 29 de Março

Resolução de Conselho de Ministros n.º 1/2006 de 2 de Janeiro

Resolução de Conselho de Ministros n.º 162/2008, de 24 de Outubro

Portaria n.º 378/94, de 16 de Junho

Portaria n.º 671/2000 de 17de Abril

Portaria n.º 95/2009, de 29 de Janeiro

WEB

http://www.dwinfocenter.org/

Sítio criado por *Larry Greenfield*, consultor da *LGI Systems*, com o objectivo de prestar ajuda sobre *Data Warehousing*.

http://www-db.stanford.edu/warehousing/warehouse.html

O projecto *WareHouse Information Prototype at Stanford* (WHIPS) tem como objectivo analisar a criação e a manutenção de DWs e desenvolver ferramentas que assegurem essas actividades.

http://www.kimballuniversity.com

Sítio da *Kimball University* com o intuito de fornecer cursos de Modelação Dimensional aplicada aos sistemas de *Data Warehousing*, com base nos livros que *Kimball* tem editado nos últimos anos.

http://datawarehouse.ittoolbox.com

Portal dedicado ao mundo do *Data Warehousing*. Contém Blogs, FAQs, *White Papers*, anúncios de empregos, etc.

http://www.inmoncif.com

Sítio da *Inmon Associates, Inc* onde é apresentada a *Corporate Information Factory* (CIF) juntamente com alguns artigos de opinião e *White Papers* sobre *Data Warehousing*.

http://www.dw-institute.com

Data Warehouse Institute é dos principais sítios de pesquisa e educação sobre *Business Intelligence* e *Data Warehousing*. Promove e patrocina conferências, seminários e cursos. Tem disponíveis no sítio *White Papers*, e publica trimestralmente o *Business Intelligence Journal*.

http://www.dm-review.com

Publicação online da *DM Review*, uma das mais conceituadas revistas de *Business Intelligence* e *Data Warehousing*. Além dos artigos publicados na revista possui artigos de opinião, *White*

Papers e portais sobre áreas especificas de *Data Waehousing*.

http://www.datawarehouse.com

Sítio desenvolvido pela *DM Review* com informações de conferências e exposições sobre *Data Warehousing*. Contém também artigos, relatórios técnicos, eventos e seminários na área de *Data Warehousing*.

http://www.tdan.com

A *Data Administration Newsletter* é uma publicação trimestral da área de administração de dados.

http://www.advisor.com

A *Advisor* é uma organização dedicada à publicação de revistas, organização de eventos, conferências e seminários sobre temas das Tecnologias da Informação e da Comunicação.

http://www.ondelette.com/OLAP/dwbib.html

Página Internet com referências bibliográficas sobre *Data Warehousing* e OLAP.

ANEXOS

Decreto-Lei n.º 280/2007, de 7 de Agosto

Portaria n.º 95/2009, de 29 de Janeiro